ACTES DE L'ATELIER IA&SANTE

1er juillet 2019 à Toulouse

Avec le soutien de l'Association française d'Informatique Médicale (AIM) et le Collège Science de l'Ingénierie des Connaissances de l'AFIA

Dans le cadre des 30èmes Journées francophones d'Ingénierie des Connaissances (IC) de la Plate-Forme Intelligence Artificielle (PFIA)

Contenu

Création automatique d'un formulaire web structuré à partir d'une ontologie : présentation du système OntoForm et de son application au cancer du sein

Fouad Sadki[1], Jacques Bouaud[2,1], Gilles Guézennec[1], et Brigitte Séroussi[1,3]

[1] Sorbonne Université, Université Paris 13, Sorbonne Paris Cité, INSERM UMR_S 1142, LIMICS, Paris, France,
[2]AP-HP, DRCI, Paris, France
[3]Hôpital Tenon, Assistance Publique – Hôpitaux de Paris, Paris, France,

(fouad.sadki.iut@gmail.com, jacques.bouaud@aphp.fr,
gilles.guezennec@univ-paris13.fr, brigitte.seroussi@aphp.fr)

Résumé : La réalisation d'une application informatique nécessite de développer une modélisation du domaine couvert par l'application, des outils de stockage en base de données et des interfaces d'acquisition et de visualisation de données. Ces développements sont souvent réalisés indépendamment les uns des autres alors que les systèmes qu'ils produisent sont intimement interconnectés. Ce manque d'interopérabilité rend ainsi couteuse la prise en compte de toute évolution applicative. OntoForm est un prototype web générant automatiquement et dynamiquement le formulaire web d'acquisition et de visualisation de données ainsi que le stockage des données à partir d'un modèle de connaissances explicite formalisé sous la forme d'une ontologie incluant modèle de données et ressources sémantiques. Ainsi, toute modification de l'ontologie est immédiatement et automatiquement répercutée sur l'ensemble des modules de l'application. Les outils générés permettent également d'interroger un système d'aide à la décision thérapeutique reposant sur la même ontologie. Cette méthode a été appliquée à la prise en charge des patientes atteintes d'un cancer du sein dans le cadre du projet européen H2020 DESIREE.

Mots-clés : OntoForm, génération de formulaire, ontologie, système d'aide à la décision, cancer du sein, stockage de données

1 Introduction

L'élaboration des bases de données permettant d'outiller le recueil, le traitement et le stockage des données suit une logique pragmatique. On commence par analyser les besoins et à formaliser les prérequis nécessaires, on conçoit ensuite le modèle des données, c'est-à-dire, comment les données vont être structurées et organisées, puis on implémente la base de données, souvent des tables relationnelles (dans le cas des bases de données relationnelles), et les outils d'acquisition de données pour tester l'ensemble (DBLC Design Stages, 2013).

Dans le cas du traitement des données par un système d'aide à la décision, la logique de construction des bases de données doit intégrer une contrainte additionnelle. En effet, et dans le cas particulier des systèmes d'aide à la décision médicale (SADM), outre les données patients qui proviennent du dossier patient informatisé (DPI), quand le SADM met en œuvre une approche à base de connaissances pour l'aide à la décision, il utilise un modèle de connaissances. Afin de permettre l'interopérabilité entre le DPI et le SADM, il faut que le modèle des données du DPI et le modèle des connaissances du SADM soient alignés. Si le

modèle des connaissances venait à changer, ce qui arrive fréquemment dans le domaine médical du fait de l'évolution continue des connaissances, il faudrait modifier le modèle des données et actualiser les alignements ce qui est fastidieux et source d'erreurs. En rendant le modèle des données et le modèle des connaissances alignés *« by design »*, si le modèle des connaissances change, alors l'ensemble du système s'adaptera automatiquement sans avoir à ajouter une surcharge de travail des développeurs, des testeurs, designers, etc.

Le projet DESIREE est un projet européen (programme H2020) dont l'objectif est de créer une plateforme web proposant plusieurs modules multidisciplinaires et collaboratifs pour la prise en charge du cancer du sein. DESIREE propose notamment une aide à la décision thérapeutique multimodale à base *(i)* de guides de bonnes pratiques, *(ii)* de l'expérience acquise par la pratique (Muro et collègues, 2017) et *(iii)* d'un raisonnement à partir de cas (Séroussi et collègues, 2018). Le module d'aide à la décision à base de guides de bonnes pratiques ou *guideline-based decision support system* (GL-DSS) utilise les données d'une patiente et les connaissances des guides de bonnes pratiques afin de proposer des recommandations de prise en charge spécifiques pour cette patiente et conformes à l'état de l'art. DESIREE utilise une interface web graphique, le DESIMS, implémentée par un des partenaires du consortium pour le recueil des données patientes. Par ailleurs, le GL-DSS se base sur un modèle de connaissances formalisé sous la forme d'une ontologie du domaine, le *breast cancer knowledge model* (BCKM) qui permet de représenter les connaissances des guides de bonnes pratiques et de raisonner sur les données des patientes. Le DESIMS et le GL-DSS qui partagent les données patientes sont ainsi interdépendants. Toute modification de l'un devra s'accompagner d'une mise à jour de l'autre, faute de quoi la chaîne de transmission de l'information de l'un à l'autre ne sera pas garantie et le GL-DSS ne pourra pas être fonctionnel, en particulier les recommandations générées (quand il y en aura car le risque principal de ce défaut d'interopérabilité est le silence du GL-DSS) ne seront pas toujours adaptées à la patiente qui sera par construction incorrectement décrite pour le GL-DSS.

Nous proposons de résoudre ce problème en générant dynamiquement et automatiquement le modèle de données et ainsi le module d'acquisition des données à partir du modèle des connaissances, c'est-à-dire l'ontologie du domaine (BCKM). Cela permet notamment de garantir que les données recueillies, sont parfaitement alignées et conformes au BCKM et donc utilisables par le GL-DSS qui pourra générer les recommandations centrées patiente appropriées, et que le stockage des données soit réalisé avec la même cohérence de format. Gonçalves a proposé d'utiliser une ontologie pour générer un outil permettant de créer un formulaire structuré destiné au recueil de données (Gonçalves et collègues, 2017). Cette approche s'inscrit a priori dans une problématique similaire à celle décrite précédemment mais l'outil généré par ces auteurs s'appuie sur un fichier de configuration XML permettant d'aligner les concepts de l'ontologie et les données du formulaire. Le problème reste donc le même, car si le modèle des connaissances change, il faudra toujours reporter ces changements dans le fichier XML et procéder à la mise à jour des alignements. Par ailleurs, cette modification convient uniquement à la partie acquisition des données, le stockage perdra toute la sémantique liée au BCKM et les données ne seront pas alignées avec l'ontologie et donc non utilisables par le GL-DSS. Stenzhorn propose également un outil d'acquisition de données en relation avec une ontologie, mais lors du stockage des données, la sémantique est de la même manière perdue (Stenzhorn et collègues, 2010). D'autres auteurs (Hochheiser et collègues, 2016) ont proposé des modèles de connaissances dans le domaine du cancer sous la forme d'une ontologie mais ils ne se sont pas intéressés aux relations de l'ontologie avec le module d'acquisition des données et au traitement des données.

OntoForm est un prototype d'une application permettant la création automatique d'un formulaire web structuré à partir d'une ontologie. Outre la création d'un formulaire web structuré, l'application offre la possibilité de stocker les données patients dans un triple store. L'application permet aussi d'éditer, supprimer, cloner, et afficher un patient et bien sûr d'afficher les recommandations suite à l'appel d'un SADM. OntoForm a été appliqué à la prise en charge des patientes atteintes d'un cancer du sein dans le cadre de l'évaluation du GL-DSS au sein de la plateforme DESIREE.

2 Matériel

2.1 Le modèle des connaissances

Le modèle des connaissances (BCKM) repose sur une ontologie qui combine à la fois *(i)* un modèle de données basé sur l'explicitation du modèle générique « entité, attribut, valeur » (EAV), et *(ii)* des ressources termino-ontologiques propres au domaine d'application, ici des notions propres au cancer du sein. Chaque élément du modèle EAV est inscrit dans l'ontologie comme une classe à part entière (Bouaud et collègues, 2018). Ces classes sont ensuite spécialisées par des concepts propres au domaine. Les relations entre entités du modèle EAV sont représentées par des *object properties*. De manière interne, un ensemble limité d'*object properties* prédéfinies organise la modélisation :

- *hasRange* définit le type de valeur de chaque attribut. Ce type peut être primitif (*Float*, *Text*, *Boolean*, etc.) ou bien hiérarchique. Dans ce dernier cas, l'ensemble des valeurs possibles est organisé par la relation de subsomption.
- *isAttributeOf* associe un attribut à une entité.
- *isRelatedTo* est la relation parente de toutes les relations entre entités du modèle EAV.

La figure 3 propose un extrait de l'ontologie utilisé dans DESIREE illustrant la modélisation du modèle EAV.

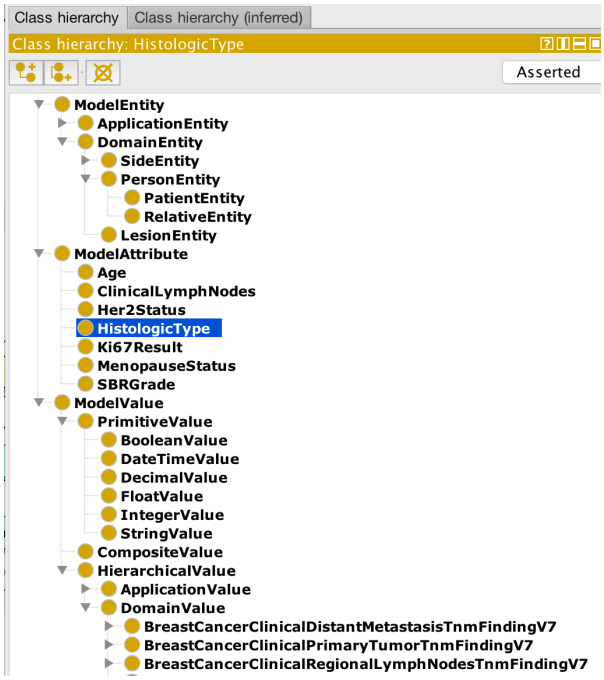


FIGURE 1 – Extrait de l'ontologie représentant le modèle EAV.

2.2 Le GL-DSSS

Le GL-DSS est un système à base de connaissances pour d'aide à la décision. Il repose sur un moteur d'inférence, Euler/EYE (Verborgh & De Roo, 2015), adapté aux technologies du web sémantique, permettant d'associer des inférences issues des relations ontologiques et des règles de production. De manière classique, le GL-DSS est alimenté par les données représentant le cas à résoudre, ici le cas d'une patiente atteinte d'un cancer du sein. La représentation du cas doit être conforme au BCKM, c'est-à-dire qu'elle utilise des notions de l'ontologie. Les bases de connaissances représentant les connaissances contenues dans les guides de bonnes pratiques

sont modélisées sous forme de règles de production dont l'expression s'appuie sur le langage NRL, inspiré de OCL (Object Constraint Language), compatible avec le modèle EAV du BCKM. Les différentes ressources, le BCKM, les règles de production, la description des cas cliniques, sont transformées en notation N3. Le format N3 est un ensemble de triplet RDF (sujet, verbe, objet). Le moteur d'inférences exécute les règles de production avec les données d'une patiente pour produire de nouveaux faits et construire les propositions de prise en charge recommandées pour cette patiente. Les recommandations issues du GL-DSS se présentent sous la forme d'un fichier XML. Le fonctionnement du GL-DSS peut être illustré en suivant l'exemple schématisé de la Figure 2.

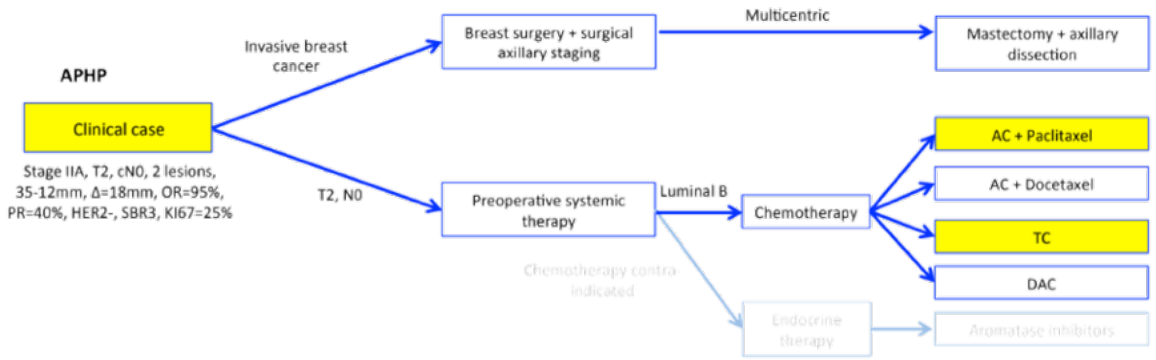


FIGURE 2 – Traitement des données permettant la génération des recommandations adaptées au cas clinique proposé en entrée. AC désigne le protocole « Doxorubicine (également appelé Adryamicine) – Cyclophosphamide », TC designe le protocole « Taxotère-Cyclophosphamide », et DAC désigne le protocole « Docetaxel-Doxorubicine-Cyclophosphamide ».

Chaque recommandation est constituée d'une ou plusieurs prescriptions ou *orders* qui décrivent des actions à réaliser. Les actions peuvent être complétées ou raffinées par d'autres recommandations et ainsi de suite (Séroussi et collègues, 2017). Par exemple, la chirurgie par tumorectomie génère une recommandation de radiothérapie du sein. Cette recommandation peut être *complétée* par la recommandation d'un « boost » sur le lit de tumorectomie. Une recommandation de chimiothérapie adjuvante peut être *raffinée* par la spécification de différents protocoles de chimiothérapie, par exemple, 4 à 6 cycles de Docetaxel - Cyclophosphamide en cas de contre-indication aux anthracyclines (cf. figure 3).

Il convient également de préciser que les recommandations issues des guides de bonnes pratiques ont des niveaux d'engagement différents. On distingue les suggestions positives, celles qui sont impératives (SHALL), recommandées (SHOULD) et possibles (MAY), et les suggestions négatives qui sont interdites (SHALLNOT), contre-indiquées (SHOULDNOT) et possibles de ne pas faire (MAYNOT). L'exemple de règle de décision proposé dans la figure 3 comporte deux *orders* qui sont recommandés (niveau d'engagement = SHOULD).

3 Méthode

L'objectif est de permettre la génération d'outils de recueil et de visualisation des données (formulaire) et de mettre en place un stockage des données qui permette l'exploitation de ces données par des outils externes, comme le GL-DSS, à partir de la spécification du modèle de données et du modèle de connaissances représentée dans l'ontologie BCKM. L'affichage des recommandations issues du GL-DSS est également un objectif.

```
Action Rule "R-APHP-Cancer-chimio-neoadj-HER2-THEN-protocoles : Si Cancer du sein ET Chimiothérapie (possible ou recommandée) ET Histologie = HER2- ET Anthra CI ALORS..."
-- ================================
        "thePatient" is a PatientEntity, "theAction" is a ActionEntity, "theOrder" represents theAction.hasOrder,
    IF
                theOrder.hasObject = thePatient
        AND     thePatient.Her2NegativePatient = true
        AND     thePatient.AvoidAnthracyclines = true
        AND     theAction.Action = SystemicChemotherapy
        AND     theOrder.Conformance is one of 'SHOULD', 'MAY'
    THEN
        [addRefinementRecoTo] theOrder from
                [build Cycles="4-6"] OrderEntity with Docetaxel to thePatient with ' ' using 'SHOULD' step 1
        AND     [build Cycles="4-6"] OrderEntity with Cyclophosphamide to thePatient with ' ' using 'SHOULD step 1
    ,
        [addComplementRecoTo] theOrder from
                [build] MessageEntity with 'Dose dense SIM or CITRON is also recommended'
    ;
```

FIGURE 3 – Exemple d'une règle de décision mettant en évidence le « raffinement » de la proposition d'une chimiothérapie, et le « complément » par une action d'affichage d'un message. Les protocoles de chimiothérapie sont recommandés (SHOULD).

OntoForm procède en plusieurs étapes principales pour la génération du formulaire à partir du BCKM : (i) extraction du modèle de données, (ii) organisation des entités sous la forme d'une arborescence, et (iii) regroupement et filtrage des attributs de chaque entité en fonction du contexte du cas patient. Sur cette base, l'élaboration des interfaces utilisateur a été réalisée. L'organisation du stockage des données est également décrite.

3.1 Extraction du modèle de données

Les entités du modèle de données sont repérées dans le BCKM comme les sous-classes de ModelEntity. Une fois les entités identifiées, pour chacune d'elles, ses attributs sont récupérés via l'*object property* isAttributeOf. Enfin, pour chaque attribut, le type de valeur est récupéré via l'*object property* hasRange.

Cette première étape est suffisante pour permettre la création d'un formulaire HTML « basique » avec un bloc pour chaque entité, puis pour chaque attribut une ligne avec le label de l'attribut suivi d'un champ de saisie dépendant du type de valeur. Par exemple, une checkbox pour un attribut booléen, un champ de saisie pour un nombre ou une chaîne de caractères, un *slider* pour un pourcentage, etc.

3.2 Structuration des entités sous la forme d'une arborescence

Les entités d'intérêt pour le formulaire sont celles pour lesquelles des relations existent entre elles dans le modèle. Aussi, toutes les *object properties* sous-classes de isRelatedTo dans le BCKM sont parcourues afin de construire un arbre de dépendance entre les entités comme illustré par la figure 4.

Cette structuration des entités permet d'organiser l'interface utilisateur, pour naviguer au sein d'entités existantes et pour permettre la création de nouvelles relations avec de nouvelles entités à partir d'une entité existante.

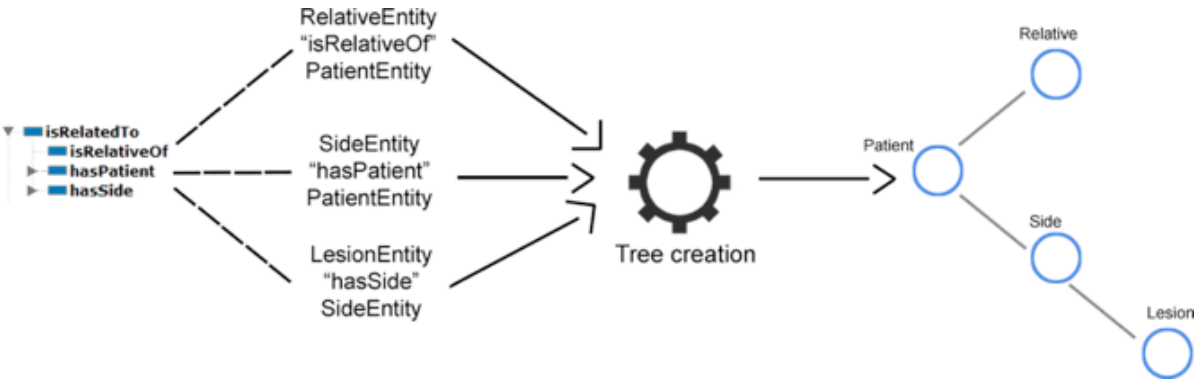


FIGURE 4 – Création de l'arbre des entités à partir de leurs relations dans le modèle.

3.3 Regroupement des attributs et filtrage en fonction du contexte

À ce stade du développement, on dispose d'une interface utilisateur fonctionnelle mais peu utilisable. En effet, on a obtenu la liste de tous les attributs par entité mais sans aucune structuration des attributs au sein des entités. Afin d'améliorer l'ergonomie de l'interface de saisie et la pertinence du formulaire pour les utilisateurs, deux propriétés ont été ajoutées aux attributs d'une même entité.

La notion de « catégorie de données » (*data categories*) permet de regrouper les attributs d'une entité en catégories en fonction de leur proximité en termes d'information pour les utilisateurs et de leur logique en termes de saisie et/ou de consultation. Ces regroupements permettent de classer les attributs dans des groupes explicites comme « Informations administratives », « Historique gynécologique », « Résultats anatomo-pathologiques », etc. Les catégories de données sont définies dans l'ontologie et attachées aux attributs en utilisant l'*object property* hasDataCategory.

Par ailleurs, la prise en charge des patientes atteintes de cancer du sein s'organise en parcours de soins. Cinq scénarios marquant les étapes de la prise en charge ont été identifiés dans le projet DESIREE (A à E). Par exemple, une patiente est en scénario A quand le diagnostic de cancer du sein vient d'être posé et qu'il n'y a eu aucune prise en charge thérapeutique. En revanche, en scenario D, la patiente a eu une chirurgie première et il s'agit de décider du traitement adjuvant. Ainsi, la pertinence de l'affichage des attributs au sein du formulaire va dépendre du scénario. Par exemple, il n'est pas approprié de renseigner la taille anatomopathologique de la tumeur (mesurée sur la pièce d'exérèse, après la chirurgie) en scénario A (avant tout traitement). Dans le BCKM, les scénarios sont définis par l'*object property* hasScenario et ces propriétés sont rattachées aux catégories de données. Ainsi, dans le scénario A (après diagnostic), certaines catégories et leurs attributs seront filtrés et n'apparaitront pas dans le formulaire, comme par exemple la catégorie « Après chirurgie » à laquelle l'attribut « Marges chirurgicales sur la composante invasive » est rattaché.

3.4 Stockage

Les données du patient recueillies via le formulaire web peuvent être stockées selon différents outils, des bases de données relationnelles ou non relationnelles comme MongoDB (MongoDB, 2019). Cependant, comme précédemment décrit, le GL-DSS prend en entrée des données d'une patiente en notation N3, et les règles de décision sont représentées en N3. De plus, l'ontologie peut aussi être exprimée dans ce format avec des triplets RDF (Caroll et collègues, 2012). Nous avons donc choisi d'utiliser des triple stores pour le stockage afin de ne pas redéfinir un schéma de base de données dans le cadre d'une base relationnelle ou bien de changer le mode de stockage pour les bases non relationnelles. De cette manière, les données seront directement sauvegardées et récupérées dans leur format N3 natif.

3.5 Implémentation

L'application OntoForm qui produit le formulaire et le stockage de données en conformité avec le BCKM est développée conformément à un modèle client/serveur. Le serveur accède à l'ontologie et aux données patientes à travers un endpoint pointant sur le triple store. L'ontologie et les données patientes sont récupérées par des requêtes SPARQL. La librairie EasyRDF (EasyRDF, 2019) a été utilisée pour formater les requêtes, les envois et les résultats. Dans la version actuelle d'OntoForm, la solution open-source « Apache Jena Fuseki » (Fuseki, 2019) est utilisée comme triple store.

Côté serveur, un contrôleur codé en PHP réceptionne les commandes du client, les traite et retourne les résultats au client après avoir fait tourner les algorithmes en interne. Le client actualise la vue en temps en réel sans recharger la page. L'ensemble est totalement asynchrone et réactif par l'usage côté client de Vue.js (Vue.js, 2019). La figure 5 montre le schéma de communication entre les différents composants.

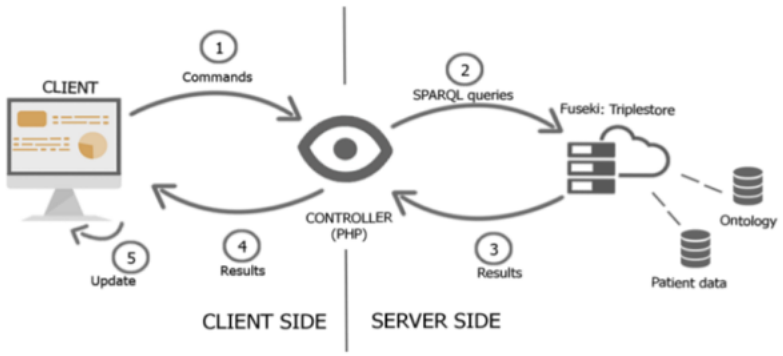


FIGURE 5 – Schéma de l'architecture générale d'OntoForm.

4 Résultats

Nous avons implémenté un module web capable de générer dynamiquement et automatiquement un formulaire structuré regroupé par entités contenant des attributs eux-mêmes regroupés en catégories à partir d'une ontologie structurée selon un modèle EAV.

Comme dans tout formulaire, chaque donnée est caractérisée par un label et un champ à compléter. Le label est récupéré du BCKM et le type du champ est interprété par la valeur de l'*object property* hasRange associée à l'attribut : le champ créé n'accepte que les valeurs numériques dans le cas d'une valeur numérique, c'est une checkbox à trois états (vrai, faux et non affecté) dans le cas d'une valeur booléenne, et dans le cas d'une valeur hiérarchique, les valeurs possibles sont récupérées du BCKM et affichées sous la forme d'un arbre au sein duquel l'utilisateur peut naviguer.

La figure 6 propose une illustration de la version actuelle du formulaire. Le bandeau à gauche de l'écran permet à l'utilisateur de gérer les entités (*patient*, *side*, *lesion*, etc.). Pour une entité sélectionnée, on affiche à droite l'ensemble de ses attributs regroupés par catégories. Lorsque l'entité « lésion » est sélectionnée, l'attribut « Type histologique » (*histologic type*) peut être renseigné par une valeur hiérarchique (menu déroulant afficahnt l'ensemble du jeu de valeurs de l'attribut à partir du BCKM). Les attributs sont regroupés par catégories comme « Histological study », « Imunohistochemistery », etc., afin de rendre la navigation plus fluide.

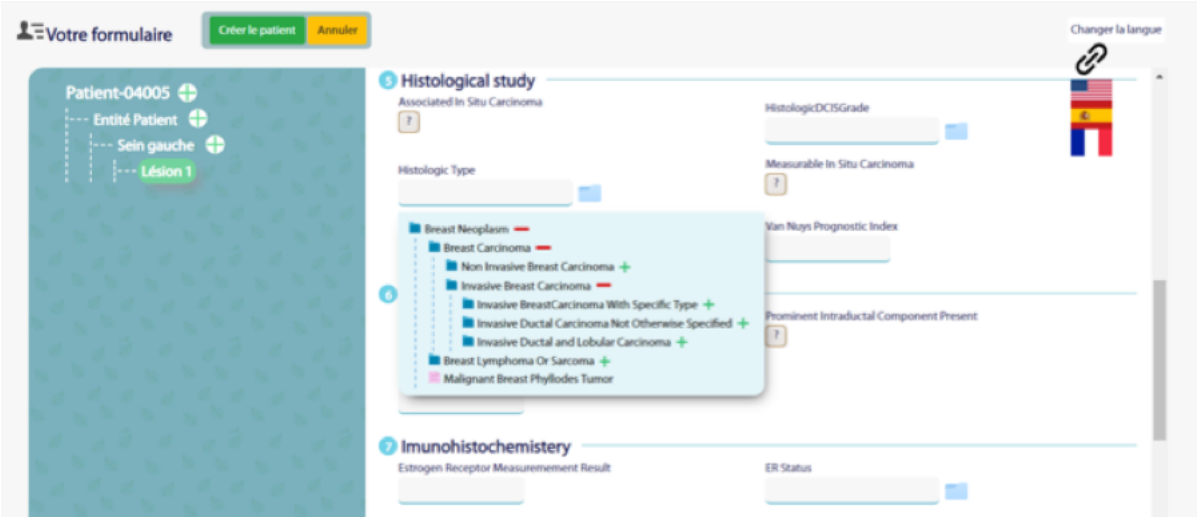


FIGURE 6 – Copie d'écran du formulaire web généré.

Une fois la patiente créée et le formulaire rempli, OntoForm peut interroger le GL-DSS et récupérer les recommandations produites au format XML. L'interprétation du XML permet l'affichage des recommandations sous la forme de parcours (cf. figure 7) où chaque ligne est une recommandation et chaque bloc correspond à une action recommandée (*order*) avec un code couleur pour le niveau d'engagement de la recommandation (vert pour SHALL jusqu'à rouge pour SHALLNOT). Une seconde option d'affichage permet de visualiser le cas clinique instancié sous la forme d'un graphe interactif tel qu'illustré par la figure 8.

FIGURE 7 – Affichage des recommandations sous la forme de séquences d'actions recommandées au sens SHOULD (en vert) et possible au sens MAY (en jaune). Les boutons bleus *Refine* permettent d'affiner l'action recommandée.

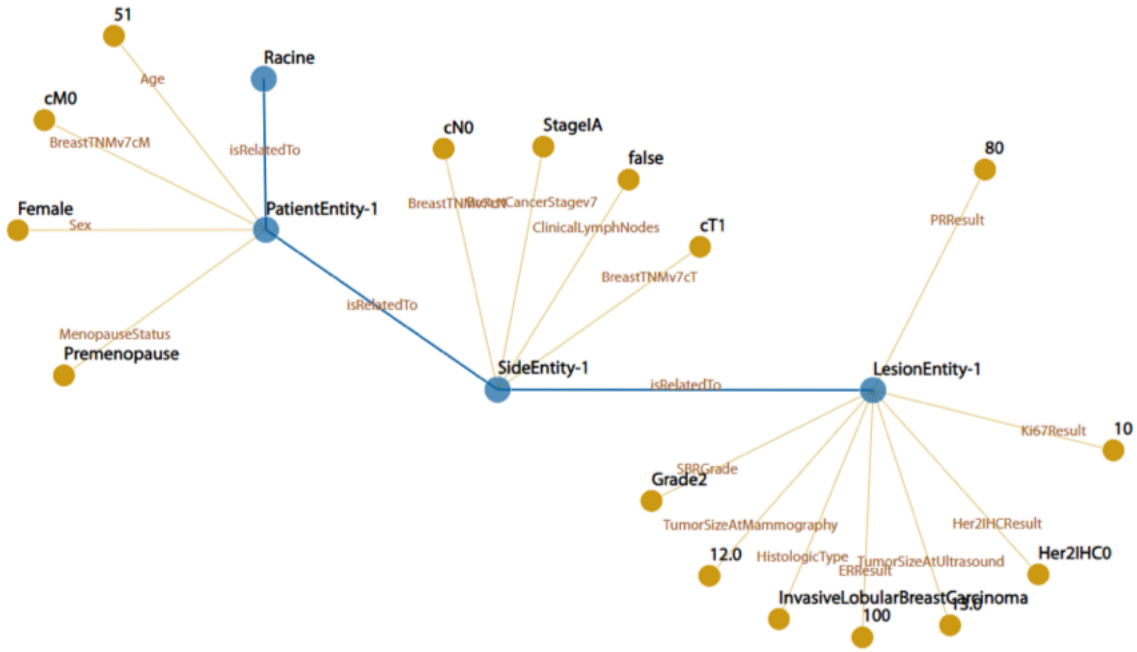


FIGURE 8 – Affichage du cas clinique instancié selon la structuration EAV du BCKM.

5 Conclusion

Nous avons développé OntoForm, un module web générant automatiquement et dynamiquement un formulaire web sémantiquement structuré. Ce formulaire basé sur une ontologie structurée par le modèle EAV est articulé autour d'entités et de catégories de données. Le stockage des données issues du formulaire est lui aussi dérivé de l'ontologie permettant ainsi l'utilisation d'outils externes comme le GL-DSS du projet DESIREE. OntoForm est auto-suffisant dans la génération du formulaire, pour le stockage et l'appel du GL-DSS.

La génération du formulaire gérant le modèle de données à partir de l'ontologie définissant le modèle de connaissances répond à la problématique principale de l'interopérabilité car si le modèle des connaissances change (BCKM dans ce cadre) le formulaire sera automatiquement actualisé et aligné par construction, sans qu'une intervention technique sur le module d'acquisition des données ne soit nécessaire. OntoForm est encore un prototype et devrait prochainement être évalué par des utilisateurs afin d'en évaluer son efficacité et ergonomie.

OntoForm a été appliqué dans le cadre du cancer du sein mais les choix d'implémentation adoptés permettent de garantir qu'avec des changements mineurs sur l'ontologie et de nouvelles règles de décision pour le GL-DSS, le système pourrait s'appliquer à d'autres domaines (issus du secteur médical ou pas). Par ailleurs, des développements additionnels d'OntoForm pourront permettre d'afficher le suivi d'un patient comme notamment l'évolution de la taille de la tumeur depuis son diagnostic par le biais de graphiques ou de variables multiples.

Références

BOUAUD J, GUÉZENNEC G, SÉROUSSI B. Combining the Generic Entity-Attribute-Value Model and Terminological Models into a Common Ontology to Enable Data Integration and Decision Support, *Stud Health Technol Inform* (2018), 541–545.

CARROLL J., HERMAN I., PATEL-SCHNEIDER P. F. OWL 2 Web Ontology Language RDF-Based Semantics (Second Edition). 2012, https://www.w3.org/2012/pdf/REC-owl2-rdf-based-semantics-20121211.pdf (dernier accès en Mai 2019).

DBLC Design Stage. Database Design. Cycle de vie d'une base de données. 2013. https://www.relationaldbdesign.com/relational-database-design/module3/dblc-design-stages.php (dernier accès en mai 2019).

EASYRDF – RDF Library for PHP. Documentation officielle d'EasyRDF. http://www.easyrdf.org/ (dernier accès en mai 2019).

FUSEKI. Apache Jena – Apache Jena Fuseki. Documentation officielle d'Apache Jena Fuseki (2019). https://jena.apache.org/documentation/fuseki2/ (dernier accès en mai 2019).

GONÇALVES R.S., TU W.S., NYULAS C.I., TIERNY M.J., MUSEN M.A. An ontology-driver tool for structured data acquisition using Web forms. J Biomed Semantics. 2017; 8: 26.

HOCHHEISER H, CASTINE M, HARRIS D, SAVOGA G, JACOBSON RS. An information model for computable cancer phenotypes. BMC Med Inform Decis Mak. 2016 Sep 15;16(1):121.

MONGODB. The most popular database for modern apps (2019). Site officiel de MongoDB. https://www.mongodb.com/ (dernier accès en mai 2019).

MURO N, LARBURU N, BOUAUD J, SÉROUSSI B, Weighting Experience-Based Decision Support on the Basis of Clinical Outcomes' Assessment. Stud Health Technol Inform. 2017;244:33-37.

SÉROUSSI B, LAMY JB, MURO N, LARBURU N, SEKAR BD, GUÉZENNEC G, BOUAUD J. Implementing Guideline-Based, Experience-Based, and Case-Based Approaches to Enrich Decision Support for the Management of Breast Cancer Patients in the DESIREE Project. Stud Health Technol Inform, 2018;255:190-194.

SÉROUSSI B, GUÉZENNEC G, LAMY JB, MURO N, LARBURU N, SEKAR BD, PREBET C, BOUAUD J. Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project. *AMIA Annu Symp Proc* 2018 Apr 16;2017.

STENZHORN H, WEILER G, BROCHHAUSEN M, SCHERA F, KRITSOTAKIS V, TSIKNAKIS M, KIEFER S, GRAF N. The ObTiMA system - ontology-based managing of clinical trials. *Stud Health Technol Inform*, 2010;160.

VERBORGH R, DE ROO J (2015). Drawing conclusions from linked data on the web: The EYE reasoner. IEEE Software. 32(3), p. 23–27.

VUE.JS. *The Progressive JavaScript Framework.* Documentation officielle de Vue.js (2019). https://vuejs.org/ (dernier accès en mai 2019).

Apport des concepts définis d'une ontologie pour l'analyse clinique de parcours de santé dans le cadre de la sclérose latérale amyotrophique

Sonia Cardoso[a,b], Xavier Aimé[b], Pierre Meneton[b], David Grabli[d], Vincent Meininger[c], Jean Charlet [b,c]

[a]Institut du Cerveau et de la moelle épinière, Paris, France,
sonia.cardoso@icm-institute.org ;
[b]INSERM, U1142, LIMICS, F-75006, Paris, France ;
Sorbonne Universités, Paris 06, UMR_S1142, LIMICS, F-75006, Paris, France ;
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S1142), F_93430, Villetaneuse, France ;
[c]Assistance publique – Hôpitaux de Paris DRCD, F-75004 Paris,, France ;
[d]Assistance Publique Hôpital Pitié Salpetrière, Département des maladies du Système Nerveux , Paris, France ;
[c] RAMSAY General de Santé, Hôpital des Peupliers, Paris, France.

Résumé : pour comprendre le parcours de santé de patients ayant une Sclérose Latérale Amyotrophique, nous devons annoter une grande quantité de données textuelles, issue d'une base de données créée par les coordinateurs du réseau SLA Île-de-France. Pour cela nous avons développé une ontologie modulaire, constituée de quatre modules, et un outil d'annotation sémantique. La spécificité de notre démarche est la création de concepts définis à différents niveaux dans les ontologies. Ces concepts définis représentent des thématiques d'intérêt pour la compréhension clinique des causes de rupture de parcours de soins. L'annotation sémantique des corpus par l'outil développé, prenant comme référentiel l'ontologie créée nous permet de réaliser secondairement un export quantitatif des concepts annotés. Cet export quantitatif de la fréquence des termes annotés et en particulier des concepts définis nous permet de réaliser des analyses statistiques permettant de tester les hypothèses cliniques émises, pour comprendre les parcours de santé des patients.

Mots-clés : Ontologie ; concepts définis ; sclérose latérale amyotrophique.

Introduction

La sclérose latérale amyotrophique (SLA) est une maladie neurodégénérative progressive des motoneurones, entrainant une paralysie progressive des muscles volontaires. Le temps de survie médian après l'apparition des symptômes est géneralement de 3 ans [1]. Cette paralysie progressive va placer les patients face à de nombreuses situations de handicap. Au-délà des aspects médicaux, les patients et leur entourage vont nécessiter différents types d'aides, (1) des aides humaines pour la réalisation d'activités de vie quotidienne comme manger, se laver, s'habiller… (2) des aides techniques de déplacements ou de communication comme un fauteuil roulant, une synthèse vocale…, mais aussi (3) un accompagnement social pour la mise en place de financement [2]. Des problèmes peuvent survenir à domicile entraînant des interruptions dans le parcours de santé et des hospitalisations. Ces ruptures peuvent avoir un impact négatif sur la santé et la qualité de vie du patient et de sa famille. Cependant, nous ne disposons d'aucune donnée sur l'origine des problèmes entraînant ces ruptures de parcours à domicile. Si les causes des interruptions de soins étaient comprises, des mesures préventives pourraient être mises en place. En France, la prise en charge des patients ayant une SLA se fait essentiellement au niveau de centre expert régional. A Paris, un réseau régional de coordination ville hôpital a été créé : le réseau de SLA Île-de-France (IdF) dont l'objectif principal est de coordonner et d'accompagner les patients, les familles, les professionnels dans la prise en charge médico-sociale. Le réseau SLA a créé une base de données textuelles recueillant (1) l'ensemble des demandes et besoins émis par les patients ou les professionnels de proximité, et (2) les actions

de coordination mises en place pour répondre aux besoins. L'hypothèse de notre travail est qu'en analysant ces corpus, il sera possible de a) comprendre les difficultés des patients et des familles à leurs domiciles, b) comprendre les actions de coordination mises en place, c) identifier les indicateurs de ruptures, ou les typologies de patient à risque.

L'exploitation de ces corpus, contenant une grande quantité d'informations, nécessite la mise en place d'outils d'ingénierie des connaissances (IC) et de traitement automatique de la langue naturelle (TALN). Nous proposons dans ce cas d'usage l'utilisation d'une ontologie modulaire et la création de concepts définis. Ces derniers serviront de variables d'intérêt lors de l'analyse statistique des données extraites des corpus, après annotation sémantique. Nous pourrons ainsi émettre des hypothèses cliniques de corrélation entre certaines variables et tenter de les vérifier. Par exemple : « en quoi le domaine de l'épuisement de l'aidant influence-t-il les composantes d'hospitalisations et les types d'actions de coordination réalisées ? »

La première section de cet article décrira l'ontologie construite, ONTOPARON, ses différents modules, ainsi que les concepts définis. La deuxième section exposera les outils d'annotation mis en place. Enfin, la troisième section exposera les premiers résultats obtenus, les perspectives et axes d'amélioration à développer.

1 Méthodes

Les spécificités de notre projet sont liées (1) au choix de modélisation des connaissances au travers d'une ontologie modulaire et (2) à la création de concepts définis ayant une dimension clinique pouvant être utilisés lors d'analyses statistiques. Le choix de la modularité s'explique pour plusieurs raisons : a) la possibilité d'utiliser secondairement une partie des modules pour l'analyse de parcours de santé dans d'autres pathologies neurodégénératives ; b) faciliter la mise à jour (manutention) des connaissances en prenant en compte les évolutions des systèmes d'aides sociales et évolutions médicales.

1.1 Création des modules de l'ontologie

Pour la création de l'ontologie la première étape a consisté en l'extraction de termes du corpus à partir d'outils founis par le TALN. Ce traitement a été réalisé en utilisant le logiciel BIOTEX[3]. Il a permis de choisir les candidats termes ayant un indicateur de fréquence important. Le corpus était constitué de 30 130 événements principaux extraits de la base de données du réseau SLA IdF, préalablement anonymisés et couvrant une période de dix ans d'activité du réseau (2005 à 2015).

Pour développer les ontologies, nous avons suivi la méthodologie générique illustrée par Charlet et al. [4] qui combine une approche descendante par l'utilisation d'une top-ontologie et une approche ascendante avec la recherche de candidats termes à partir des corpus. Cette méthode permet d'accéder aux termes représentant les concepts utilisés. Chaque concept d'ONTOPARON est désigné par un terme préféré en anglais et en français, ainsi que par des termes alternatifs (termes synonymes, acronymes, abbréviations prenant en compte les spécificités orthographiques liées au contexte de coordination). Les modules de l'ontologie ont été créées sous Protégé[1]. Il n'existe pas de consensus inter-coordinateurs du réseau SLA IdF, pour l'utilisation d'abbréviations communes. Cette diversité d'usage a nécessité le recueil de l'ensemble des termes et abbrévations utilisés pour un même concept (par exemple, le concept « médecin traitant » a huit termes alternatifs pouvant le dénoter : med tt, mt, médecin de famille, med ttt, méd t, méd tt, médecin généraliste, généraliste). L'approche descendante s'est faite lors de la construction des modules par la prise en compte des spécificités des corpus et notamment la prise en compte des concepts issus du domaine social francais. En effet il n'existe pas d'ontologie contenant de façon spécifique l'ensemble des concepts liés aux prestations et organismes sociaux délivrant les allocations sociales (comme les Maisons départemenatles des

[1] https://protege.stanford.edu

personnes handicapés ou les prestations de compensation du handicap), qui sont des élements importants dans le suivi de patient ayant une SLA. La seconde étape de ce travail fut l'alignement des concepts et l'enrichissement de l'ontologie. Pour cela nous avons utilisé l'outil HeTOP [5] afin d'aligner les concepts de nos modules ontologiques avec d'autres terminologies en utilisant leurs codes UMLS.

L'analyse des candidats termes nous a permis de définir quatre dimensions principales : (1) une dimension contenant les concepts géneriques communs à tous les champs, (2) une dimension médicale liée à la pathologie, (3) une dimension socio-environnementale liée à l'environnement familial et géographique du patient et (4) une dimension de coordination liée aux activités entreprises par les coordinateurs du réseau SLA IdF. Si, initialement, nous avons débuté par la construction d'une ontologie monolithique, rapidement nous nous sommes orientés vers la construction d'une ontologie modulaire, constituée de quatre modules de domaine et d'un module de consolidation. Une ontologie modulaire est définie comme un ensemble de modules qui sont des « composants réutilisables d'une ontologie plus grande ou plus complexe, qui est autonome mais qui présente une association définie avec d'autres modules d'ontologie, y compris l'ontologie originale » [6].

Le premier module est le module noyau. Il contient l'ensemble des concepts de haut niveau commun aux trois ontologies, comme les objets idéaux, les agents, les processus, les modalités. Cette ontologie s'est inspirée du travail mené par Charlet *et al.* sur l'ontologie MENELAS [7]. Ce module contient également l'ensemble des relations liant les concepts de haut niveau, et les concepts définis.

Le module médical possède un niveau plus spécifique par rapport à l'ontologie noyau. On y trouve les agents médicaux (médecin, neurologue, kinésithérapeute...), les processus médicaux (consultation, hospitalisations...), les objets médicaux (les médicaments, les sondes...)… On y trouvera également les structures anatomiques, les suppléances respiratoires mises en place dans le cadre de la SLA (la gastrostomie, la trachéotomie...). Il regroupe l'ensemble des concepts liés directement à la pathologie et à la prise en charge médicale.

Le module socio-environnemental rassemble l'ensemble des concepts liés à la vie de la personne dans son environement familial et social. Sont modélisés les agents sociaux (la famille, les auxiliaires de vie...), les actions sociales (la demande d'allocation, l'épuisement de l'aidant...), les objets sociaux (aides techniques, carte vitale...). Nous avons utilisé le travail mené sur Ontopsychia [8] pour ré-utiliser certains concepts de l'ontologie.

Le module coordination est composé en grande partie par les missions spécifiques de coordination. Ces actions de coordinations sont de plusieurs types : des actions de communication, des actions d'évaluation des besoins, des actions de recherche de ressource. Pour cette modélisation, nous nous sommes inspirés en partie des travaux de Popejoy *et al.* [9] qui ont créé la Nursing Care Coordination Ontology.

Le dernier module dénommé ONTOPARON, est le module de consolidation. Il permet de regrouper l'ensemble des quatre modules précédents. Une fois les différents modules importés, nous utilisons les fonctions de raisonnement et d'export à partir de protégé, pour obtenir une ontologie complète qui sera implémentée dans l'outil d'annotation sémantique développé au sein du laboratoire. Des informations plus précises sur les motivations et la réalisation modulaire de l'ontologie[2] sont disponibles dans [10].

1.2 Création des concepts définis

L'organisation des connaissances sous forme d'ontologie, nous permet de créer des concepts définis (*equivalent class*). Lors de l'utilisation du raisonneur l'ensemble des concepts liés par une relation seront inférés sous le concept défini. Nous souhaitons comprendre s'il existe des corrélations entre différentes variables, dans le cadre de la coordination de parcours de santé pouvant expliciter les ruptures de parcours de santé des patients. Le fonctionnement du réseau SLA est basé sur la sollicitation du réseau, par les agents (patient, entourage, famille,

[2] L'ontologie est accessible à https://bioportal.bioontology.org/ontologies/ONTOPARON

professionnel de proximité…) lorsqu'ils en ont besoin. Chacune de ces sollicitations va générer un ou plusieurs évènements en fonction des actions qui devront être mises en place par les coordinateurs. L'objectif de notre travail est de comprendre quels sont les éléments intervenant dans les parcours de santé des patients. Existe-t-il une typologie de patients plus à risque de rupture ? Certains patients ont-ils plus d'hospitalisations que d'autres, y a-t-il des facteurs sociaux associés ? Quelles actions de coordination sont mises en place, existe-t-il des disparités?

Notre hypothèse est que l'annotation sémantique de la base évènementielle SLA, par l'outil d'annotation OnBaSAM utilisant comme référence l'ontologie ONTOPARON peut nous apporter de nombreux éléments de compréhension des parcours. Nous souhaitons savoir si la présence de certaines variables comme « l'épuisement » ou la « douleur », ont une incidence sur les parcours de santé des patients ? Quelle est la part d'action de coordination réalisée dans le domaine social ou dans le domaine médical ? Pour tenter de répondre à ces questions nous avons fait le choix de créer des concepts définis. Leur fonction est de regrouper l'ensemble des concepts présents dans les différents modules et appartenant à une thématique commune. Ces concepts définis sont construits et créés comme étant des variables d'intérêt d'analyse clinique des parcours de santé, leur fréquence sera un indicateur des « problématiques » rencontrées par les patients. La figure 1 montre les concepts définis de haut niveau. D'autres concepts définis de plus bas niveaux sont présents dans les différents modules comme : (1) ActionDeCoordination dans le module coordination, qui regroupe l'ensemble des actions de guidage, de communication, de recherche de ressources réalisées par les coordinateurs. (2) Dans le module social nous avons créé le concept défini ProcessusSocial qui regroupe l'ensemble des demandes et actions faites dans le domaine social. (3) Dans le module médical nous avons créé le concept défini EtatCognitif qui regroupe l'ensemble des termes faisant référence à des signes cliniques, symptômes et diagnostics liés à une altération cognitive du patient.

La structure de l'ontologie ne permet pas toujours d'avoir une vision globale, alors qu'un processus et un état se trouvent à deux niveaux différents dans l'ontologie, la création d'un concept défini à un niveau supérieur permet de lier les concepts issus de ces deux axiomes, qui d'un point de vue clinique vont évoquer une même problématique clinique. Ex : l'épouse est épuisée (état-social), alerte d'un professionnel sur l'épuisement de l'épouse (action sociale) et recherche d'un séjour de répit (action de coordination). Ces trois concepts sont réunis par la restriction « aPourThematique some ThematiqueEpuisement » sous le concept défini DomaineEpuisement. Dans cet exemple, après raisonnement, le concept défini DomaineEpuisement subsumera l'ensemble des concepts portant la restriction sus-nommée ; ces concepts pouvant se trouver dans les différents modules. Lors de l'annotation des corpus et l'exportation de la fréquence des termes annotés, nous pourrons réaliser secondairement une analyse statistique pour rechercher les corrélations existantes entre différentes variables d'intérêt. Au total 33 concepts définis furent créés dans les différents modules ontologiques.

2 Annotateur OnBaSAM

Au sein du laboratoire nous développons un annotateur sémantique basé sur une plateforme GATE (General Architecture for Text Engineering) qui utilise l'ontologie ONTOPARON comme référentiel sémantique. Différents pipelines ont été créés, permettant plusieurs niveaux de traitement des corpus : (1) Pré-traitement par la normalisation et tokenization, découpage en phrases, application de la lemmatization (TreeTagger) et *Pos tagging* afin de reconnaitre la catégorie grammaticale. Ce pipeline permet de réaliser la correction orthographique des corpus, augmentant l'amélioration du repérage des concepts, lors de l'annotation. (2) Le second pipeline permet l'annotation des entités nommées par le repérage des concepts de l'ontologie par les preflLabel, les altLabel du concept et le repérage des instances. (3) L'option d'export des annotations effectuées vers un tableur a été implémentée. Pour chaque patient, est associé pour concept le nombre d'occurrences repérées par OnBaSAM. Cet export peut être total en

prenant en compte l'ensemble des concepts de l'ontologie (2386) ou bien spécifique en ne retenant que les variables d'intérêt, qui sont dans notre cas les concepts définis (33).

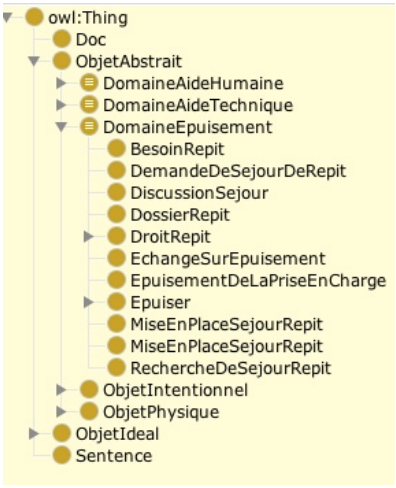


Figure 1- Vue de l'ontologie Noyau contenant les concepts définis de haut niveaux.

3 Résultats

Nous avons réalisé l'annotation sémantique en utilisant l'outil OnBaSAM sur 611 dossiers de patients (décédés) afin d'avoir des parcours de santé complets, cela a représenté 28 052 évènements primaires et secondaires. Ces patients ont été inclus entre 2013 et 2017. Le nombre d'évènements pour chaque patient varie de 1 à 188, avec une médiane de 23 et une moyenne de 31,6 évènements par patient. Nous avons souhaité répartir les patients en groupe homogène, en fonction du nombre d'évènements. Le nombre d'événement indique si les patients et les familles sollicitent le réseau SLA IdF. En utilisant une répartition logarithmique nous obtenons trois groupes de patients comme indiqué dans la table 2. Nous avons utilisé le logiciel JMP[3] afin de réaliser les différents traitements statistiques des données.

Table 2 - Répartition des patients par nombre d'évènements.

	T1	*T2*	*T3*
Nombre d'évènements	*1 à 13*	*14 à 32*	*33 à 188*
Nombre de patients	*171*	*216*	*224*

Afin d'illustrer l'apport et l'intérêt des concepts définis dans notre projet, nous allons prendre l'exemple du concept défini « DomaineEpuisement », ce concept défini regroupe l'ensemble des concepts faisant référence à l'épuisement de l'aidant (par exemple : mise en place d'un séjour de répit, alerte sur épuisement…) dans les différents modules ontologiques. Nous exposerons quelque exemples de corrélations existantes entre ce concept et d'autres variables d'intérêt et expliciterons leurs apports dans l'analyse clinique des parcours.

La première question fut de savoir si le domaine de l'épuisement était présent de manière similaire dans les différents groupes. La figure 2 illustre cette répartition. Il semble que le groupe de patients T3, ayant le plus grand nombre d'évènements, soit aussi le groupe sur lequel la notion d'épuisement est la plus fréquente, la présence de l'épuisement de l'aidant peut donc générer plus d'évè&nements et peut être plus d'actions de coordination.

[3] https://www.jmp.com/fr_fr/home.html

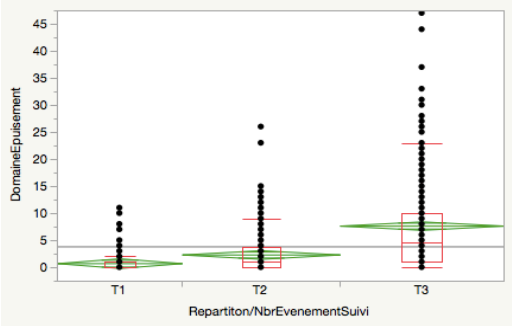


Figure 2 – Analyse de la variance du DomaineEpuisement en fonction de la répartition des patients par nombre d'évènements.

Nous avons souhaité savoir secondairement si la présence de DomaineEpuisement pouvait être lié, comme l'indique la littérature [11], à la présence d'une altération de l'état cognitif du patient. La présence de troubles cognitifs est référencée dans la littérature sur la SLA (trouble du comportement, présence d'une démence fronto-temporale (DFT)) [12]. Pour tester cette hypothèse nous avons réalisé un test statistique pour évaluer le degré de corrélation entre ces deux variables que sont les concepts définis DomaineEpuisement et EtatCognitif. La valeur de la p-value, obtenue, entre ces deux variables est de $p < 0{,}0001$ et le coefficient de corrélation est de $r = 0{,}23$. On peut donc penser qu'il existe un lien entre la présence de troubles cognitifs chez le patient et l'apparition de phénomène d'épuisement chez l'aidant. Ces résultats vont dans le sens d'études menées sur l'épuisement des aidants dans la maladie d'Alzheimer [13].

La corrélation faible entre l'épuisement de l'aidant et la présence de trouble cognitif trouvé, nous avons souhaité savoir si l'âge du patient était une variable intervenant dans l'apparition de l'épuisement. Les résultats de l'analyse de corrélation entre ces deux variables, a montré avec une certaine certitude (p-value = 0,0079) que l'âge n'intervient pas dans le DomaineEpuisement.

Nous avons ensuite fait l'hypothèse que la présence de l'épuisement pouvait intervenir sur le nombre d'actions de coordination, probablement pour mettre en place des aides humaines, des aides techniques ou des solutions de séjour de répit afin de soulager l'aidant.

Nous avons testé les deux variables du DomaineEpuisement et ActionDeCoordination, la valeur obtenue est une p-value <0,0001 avec un $r = 0{,}55$ indiquant une faible corrélation entre la présence de l'épuisement et la mise en place d'action de coordination comme indiqué à la figure 4.

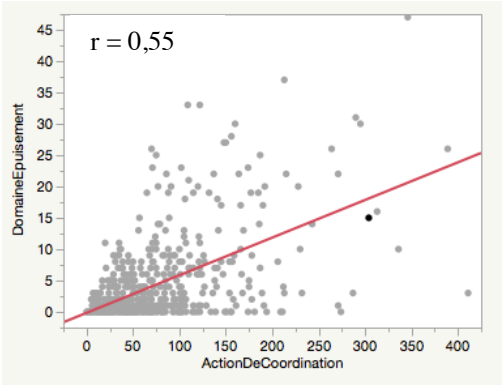


Figure 4 - Analyse bivarié de DomaineEpuisement par ActionDeCoordination.

Ces résultats sont les prémices d'analyses statistiques qui devront être développées plus amplement dans notre projet. Ces premiers résultats nous permettent d'envisager la mise en

place d'action de prévention en coordination. La présence, par exemple, des premiers signes d'altération cognitive peut inciter les coordinatrices à planifier des appels de suivi afin d'évaluer l'épuisement de l'aidant, de faire un point sur les ressources et dispositifs mis en place.

4 Discussion

Différents outils furent créés et utilisés pour l'exploitation des corpus de coordination, afin de comprendre les parcours de santé des patients ayant une SLA, mais il reste encore de nombreux points à mettre en place. Les premiers résultats obtenus en exploitant le potentiel des concepts définis sont encourageants, cependant de nombreux axes de travail doivent encore être développés pour améliorer et évaluer les performances des différents outils. Deux dimensions doivent être poursuivies, la première concerne la dimension de développement informatique, et la seconde dimension est l'analyse des résultats utilisant des méthodologies rigoureuses de statistiques associées à une expertise clinique.

Pour la dimension informatique nous souhaitons (1) dans un premier temps poursuivre le travail mené avec les experts du domaine afin de valider totalement les modules ontologiques. (2) Secondairement nous souhaitons évaluer notre outil d'annotation (OnBaSAM). Pour cela nous développons actuellement l'outil PRONTO qui nous permettra, de faire valider par les experts du domaine les annotations faites par le système, un module intégré nous permettra d'obtenir directement les valeurs de Rappel, Précision et F-mesure. Ces résultats seront comparés à des systèmes équivalents comme l'ECMT, pour lequel des évaluations ont déjà été faites. (3) Le dernier grand axe est de poursuivre l'amélioration de la détection de la négation et de l'hypothèse dans l'annotation des corpus. Notre hypothèse est que, au vu de notre corpus, l'amélioration de la détection de la négation et du conditionnel dans les corpus, ne devrait pas modifier considérablement les valeurs exportées.

Pour l'analyse des résultats, nous devons allier l'expertise de professionnels cliniques (médecins, coordinateurs) et de professionnels des statistiques. Ce partenariat nous permettra de construire un modèle fiable, permettant de valider ou invalider les hypothèses de recherche. L'objectif étant de « trouver » des indicateurs pouvant expliciter les ruptures de parcours dans la SLA. Si l'ensemble du modèle est validé dans notre cas d'usage, les outils pourront être testés sur d'autres corpus de coordination pour d'autres maladies neurodégénératives.

5 Conclusion

Dans cet article nous avons souhaité présenter, notre démarche utilisant les capacités de raisonnement et d'inférences possibles dans les ontologies, pour créer des concepts définis qui, au-delà de l'aspect sémantique, ont une réelle dimension et expression clinique. La modularisation des ontologies et l'association d'outils de traitement automatique de la langue naturelle, que nous avons développés, nous permettent d'annoter des corpus et d'en extraire les connaissances pour tenter de comprendre les éléments intervenant dans les parcours de santé de patients SLA.

Références

[1] P. Couratier, P. Corcia, G. Lautrette, M. Nicol, P.-M. Preux, and B. Marin, Epidemiology of amyotrophic lateral sclerosis: A review of literature, *Revue Neurologique*. **172** (2016) 37–45. doi:10.1016/j.neurol.2015.11.002.

[2] M.-H. Soriani, and C. Desnuelle, Care management in amyotrophic lateral sclerosis, *Revue Neurologique*. **173** (2017) 288–299. doi:10.1016/j.neurol.2017.03.031.

[3] J.A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, BioTex: A system for biomedical terminology extraction, ranking, and validation, in: SIMBig: Symposium on Information Management and Big Data, Cusco, Peru, 2014. https://hal.archives-ouvertes.fr/hal-01136531 (accessed June 19, 2019).
[4] J. Charlet, B. Bachimont, and M.-C. Jaulent, Building medical ontologies by terminology extraction from texts: An experiment for the intensive care units, *Computers in Biology and Medicine*. **36** (2006) 857–870. doi:10.1016/j.compbiomed.2005.04.012.
[5] J. Grosjean, L.F. Soualmia, K. Bouarech, C. Jonquet, and S.J. Darmoni, An approach to compare bio-ontologies portals, *Stud Health Technol Inform*. **205** (2014) 1008–1012.
[6] J. Pathak, T.M. Johnson, and C.G. Chute, Survey of modular ontology techniques and their applications in the biomedical domain, *Integr Comput Aided Eng*. **16** (2009) 225–242. doi:10.3233/ICA-2009-0315.
[7] J. Charlet, B. Bachimont, L. Mazuel, F. Dhombres, M.-C. Jaulent, and J. Bouaud, OntoMénélas. Motivations et retours d'expérience sur l'élaboration d'une ontologie noyau de la médecine, *Techniques et sciences informatiques*. **31** (2012) 125–147. doi:10.3166/tsi.31.125-147.
[8] R. Marion, A. Xavier, K. Marie-Odile, and C. Jean, Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts, *Studies in Health Technology and Informatics*. (2015) 221–223. doi:10.3233/978-1-61499-512-8-221.
[9] L.L. Popejoy, M.A. Khalilia, M. Popescu, C. Galambos, V. Lyons, M. Rantz, L. Hicks, and F. Stetzer, Quantifying care coordination using natural language processing and domain-specific ontology, *J Am Med Inform Assoc*. **22** (2015) e93–e103. doi:10.1136/amiajnl-2014-002702.
[10] S. Cardoso, X. Aime, V. Meininger, D. Grabli, K.B. Cohen, and J. Charlet, De l'intérêt des ontologies modulaires. Application à la modélisation de la prise en charge de la SLA, in: S. Ranwez (Ed.), 29es Journées Francophones d'Ingénierie Des Connaissances, IC 2018, AFIA, Nancy, France, 2018: pp. 121–128. https://hal.archives-ouvertes.fr/hal-01839571 (accessed May 14, 2019).
[11] J. Oh, and J.A. Kim, Factor analysis of the Zarit Burden Interview in family caregivers of patients with amyotrophic lateral sclerosis, *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*. **19** (2018) 50–56. doi:10.1080/21678421.2017.1385636.
[12] P. Couratier, B. Marin, G. Lautrette, M. Nicol, and P.-M. Preux, Épidémiologie, spectre clinique de la SLA et diagnostics différentiels, *La Presse Médicale*. **43** (2014) 538–548. doi:10.1016/j.lpm.2014.02.013.
[13] H. Amieva, L. Rullier, J. Bouisson, J.-F. Dartigues, O. Dubois, and R. Salamon, Attentes et besoins des aidants de personnes souffrant de maladie d'Alzheimer, *Revue d'Épidémiologie et de Santé Publique*. **60** (2012) 231–238. doi:10.1016/j.respe.2011.12.136.

Processus d'intégration de ressources termino-ontologiques en santé

Jean Noël Nikiema[1], Vianney Jouhet[1,2], Fleur Mougin[1]

[1] UNIVERSITÉ DE BORDEAUX, INSERM UMR 1219, Centre de recherche Bordeaux Population Health, équipe ERIAS, Bordeaux, France
prenom.nom@u-bordeaux.fr

[2] CHU DE BORDEAUX, Pôle de santé publique, Service d'information médicale, Bordeaux, France

Résumé : Notre travail illustre le besoin d'intégrer des ressources termino-ontologiques (RTOs) hétérogènes en santé. Dans ce cadre, nous avons étudié trois grands types de processus : l'alignement, l'intégration et le cross-linking. Plus précisément, nous avons réalisé l'alignement de RTOs décrivant des analyses biologiques, l'intégration de RTOs liées aux médicaments et le cross-linking de RTOs diagnostiques en cancérologie. L'implémentation de chacun de ces processus a permis de mettre en évidence les intérêts d'utiliser une RTO de support. Premièrement, cela a résolu des conflits sémantiques en procédant au filtrage de mappings erronés et à la désambiguïsation de mappings multiples. Deuxièmement, nous avons pu trouver automatiquement des relations transversales entre des entités différentes mais complémentaires (objectif du cross-linking). Enfin, l'utilisation d'une RTO de support a permis d'évaluer indirectement la qualité des sources de connaissances impliquées dans le processus mis en jeu.

Mots-clés : Alignement, intégration sémantique, cross-linking.

1 Contexte

Dans le domaine de la santé, il existe un nombre très important de sources de connaissances (Joubert *et al.*, 2009), qui vont de simples terminologies, classifications et vocabulaires contrôlés à des représentations très formelles que sont les ontologies (Studer *et al.*, 1998). Nous utilisons par la suite le terme de ressource termino-ontologique (RTO) pour désigner ces différentes sources de connaissances (Bourigault *et al.*, 2004).

Les RTOs biomédicales constituent un groupe hétérogène puisqu'elles ont été créées avec des niveaux de complexité différents. L'hétérogénéité de ce groupe rend leur interopérabilité complexe (Merabti *et al.*, 2009). En effet, l'utilisation secondaire des données de santé (Meystre *et al.*, 2017) pour la recherche, la définition de politiques de santé et la médecine personnalisée (Garcia *et al.*, 2013) sont autant de champs nécessitant l'intégration de données de natures diverses, provenant de différents systèmes d'information et codées suivant des RTOs différentes. Il est ainsi nécessaire de pouvoir utiliser ces RTOs de manière conjointe en ayant une vue complète et cohérente.

Dans la littérature, trois grands types d'hétérogénéité entre RTOs ont été recensés (Da Silva *et al.*, 2006). On distingue ainsi :

— l'hétérogénéité **syntaxique** : elle correspond aux différences dues au langage utilisé pour décrire les RTOs. Il s'agit de différences dans les formats d'écriture des RTOs (Resource Description Framework (RDF [1]), Simple Knowledge Organization System (SKOS [2]), Web Ontology Language (OWL [3]), etc.),

— l'hétérogénéité **structurelle** : elle correspond aux différentes manières de représenter des données dans un même format (modèle d'écriture des termes, type d'organisation hiérarchique des notions, etc.),

— l'hétérogénéité **sémantique** : elle correspond aux différences dans les notions représentées (maladies, processus biologiques, actes médicaux, actes infirmiers, etc.).

1. https ://www.w3.org/RDF/
2. https ://www.w3.org/TR/skos-reference/
3. https ://www.w3.org/TR/owl-features/

Cette étude décrit l'intérêt d'utiliser une RTO de support pour garantir l'interopérabilité sémantique entre RTOs. La première section présente les différents processus applicables pour surmonter les hétérogénéités entre RTOs. Les trois sections suivantes introduisent les techniques que nous avons implémentées pour mettre en œuvre les processus précédemment identifiés. Nous discutons finalement l'intérêt d'utiliser une RTO de support tel que cela a été mis en évidence dans chaque processus.

2 Cadre d'étude

Les correspondances entre entités de deux RTOs peuvent être retrouvées manuellement par des experts du domaine. Cependant, les RTOs pouvant contenir une grande quantité d'entités, il s'agit d'un processus qui peut être long et fastidieux. L'alternative est d'établir les correspondances de manière automatique par la création de **mappings**, qui consiste à déterminer une expression formelle de la relation sémantique entre deux entités. Un mapping est souvent représenté par un quintuplet <id, e_1, e_2, r, n>, où e_1 et e_2 sont les deux entités à lier, *id* est l'identifiant de la correspondance, *r* la relation sémantique entre les deux entités et *n* la mesure de confiance associée (Euzenat *et al.*, 2011). Deux processus existent pour établir des correspondances entre des RTOs. (1) L'**alignement** qui vise à créer des mappings entre les entités des différentes RTOs (Euzenat *et al.*, 2011). (2) L'**intégration** qui consiste à créer une nouvelle RTO en utilisant des RTOs préexistantes (Pinto *et al.*, 1999). En pratique, cela nécessite d'établir des mappings entre les entités des RTOs à intégrer, puis à réorganiser les RTOs dans une structure conjointe et unique.

Dans la littérature, ce sont essentiellement les relations d'**équivalence** et de **subsomption** qui sont trouvées lors de l'identification des mappings. Les entités représentant des notions différentes sont au mieux associées via des relations de **disjonction**. Ainsi, quand les notions entre les RTOs à relier sont distinctes mais complémentaires, les solutions proposées sont insuffisantes. Les travaux existants se contentent d'organiser de manière cohérente les entités représentant des notions différentes mais ne créent pas de liens directs entre les entités en cas de complémentarité. Pour répondre à cette problématique, nous introduisons un autre processus : le **cross-linking**. Ce processus vise à créer des mappings (relations d'équivalence ou de subsomption), à organiser de manière cohérente les entités, puis à identifier des relations transversales entre les entités qui sont différentes mais complémentaires (*e.g.*, mappings entre gènes et médicaments, entre maladies et données géographiques). Il repose sur deux étapes : (1) l'ancrage à une RTO de support, et (2) la dérivation suivant la RTO de support.

3 Processus d'alignement : cas des analyses biologiques

La LOINC® (Logical Observation Identifiers Names and Codes (Bodenreider *et al.*, 2018)) est une RTO de référence dans le domaine des analyses biologiques. Procéder à l'alignement de RTOs locales avec la LOINC permet d'assurer l'utilisation conjointe des données de biologie provenant de plusieurs structures de soins. Ainsi, en alignant les entités des RTOs locales avec celles de la LOINC, les entités LOINC peuvent être utilisées pour obtenir des données comparables à travers différents systèmes d'information.

Nous avons aligné la RTO locale du CHU de Bordeaux à la LOINC. La méthodologie reposait sur trois étapes. La première étape a consisté au pré-traitement des libellés de la RTO locale (Nikiema *et al.*, 2017b). La deuxième étape visait à calculer la similarité morphosyntaxique entre les tokens constitutifs des libellés de la RTO locale et de la LOINC. Dans la troisième étape, nous avons utilisé la structure de la LOINC pour procéder au filtrage des mappings obtenus (Figure 1). En effet, les éléments définitionnels des concepts de la LOINC sont délimités dans chacun des libellés grâce à une ponctuation précise. Le caractère ":" sépare les concepts LOINC en leurs éléments principaux, comme suit :
<composant/analyte>:<propriété>:<temps>:<milieu biologique>:<échelle>:<méthode>.
Ainsi, comme déjà implémenté dans (Mary *et al.*, 2017), nous avons créé des relations entre le concept LOINC et chacun de ses éléments définitionnels. Chaque relation a été nommée en

combinant le préfixe *has_* et le type d'élément définitionnel (*has_component*, *has_property*, etc.).

Nous avons utilisé ServoMap (Diallo, 2014) pour créer des mappings entre les concepts TLAB et les éléments définitionnels des concepts LOINC. Ensuite, des relations de mapping ont été créées entre les concepts de la LOINC et ceux de la RTO locale qui partageaient le même analyte. Ces mappings ont ensuite été filtrés grâce à la structure de la LOINC par la suppression de mappings erronés, à savoir des mappings entre concepts n'appartenant pas au même chapitre et ne décrivant pas le même milieu biologique ou la même méthode. Dans ce processus d'alignement, la structure de la LOINC a donc permis de procéder à la correction de mappings, palliant ainsi les limites de la RTO locale.

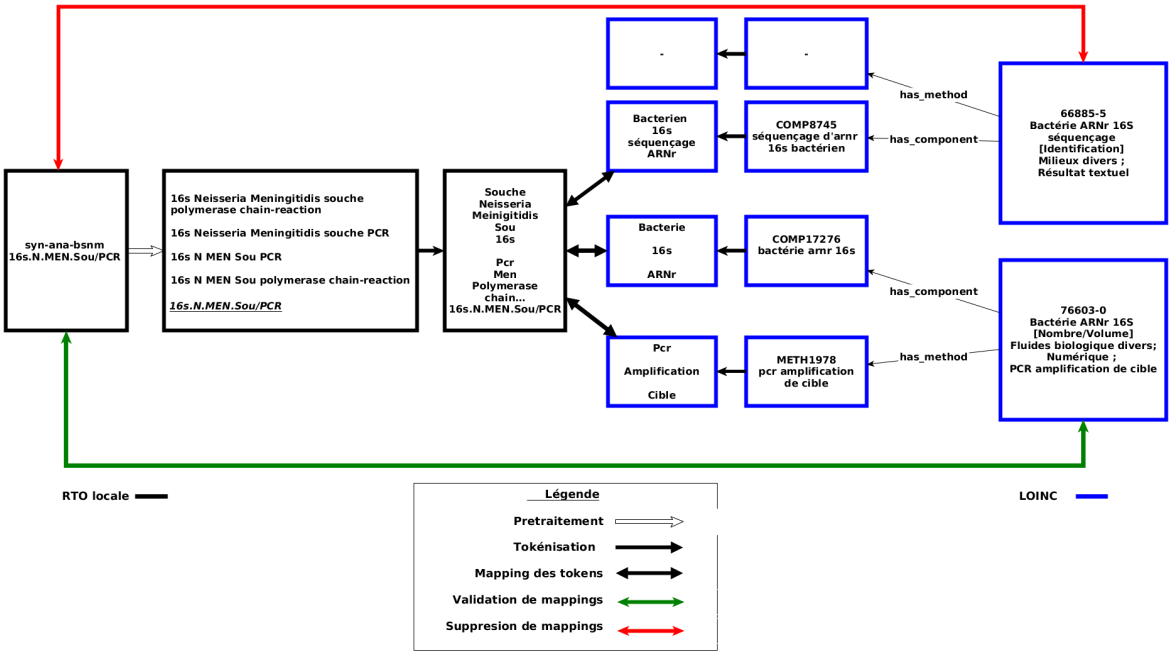


FIGURE *1 – Exemple d'alignement d'un concept de la RTO locale du CHU de Bordeaux à un concept LOINC.*

4 Processus d'intégration : cas de la représentation du médicament

Dans un deuxième travail, nous avons intégré RxNorm (Bodenreider *et al.*, 2018) à la sous-partie de la SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) (Bodenreider & James, 2018; Bodenreider *et al.*, 2018) décrivant les connaissances sur les médicaments afin d'alimenter la SNOMED CT avec les entités de RxNorm. La SNOMED CT étant une référence au niveau international et RxNorm étant utilisée aux États Unis, procéder à l'intégration de ces deux RTOs vise à rendre interopérables les données sur le médicament présentes dans les systèmes d'information sanitaire (SIS) aux États Unis à n'importe quel SIS dans le monde qui utilise la SNOMED CT. De plus, un nouveau modèle de représentation du médicament a été décrit au sein de la SNOMED CT (Bodenreider & James, 2018). Ce modèle étant basé sur les recommandations internationales regroupées au sein de l'IDMP (Identification of Medicinal Products) (Agency, 2016), l'intégration permettra d'évaluer la conformité de RxNorm aux règles internationales de description du médicament.

Nous avons représenté les concepts RxNorm selon le modèle de description du médicament utilisé par la SNOMED CT (Bodenreider & James, 2018). Ainsi, les médicaments dans

RxNorm ont été décrits en OWL grâce à leurs éléments définitionnels (substance, unité de mesure, dose, etc.). Nous avons ensuite fusionné cette représentation de RxNorm à la structure de la SNOMED CT, résultant en une nouvelle RTO composée de RxNorm et de la SNOMED CT. Au sein de cette nouvelle RTO, nous avons créé des mappings d'équivalence entre les éléments définitionnels des entités de RxNorm et celles de la SNOMED CT. Enfin, nous avons généré la structure inférée de cette nouvelle RTO en utilisant le raisonneur ELK (Abburu, 2012). Ce choix est motivé par le fait que ce raisonneur a été décrit comme étant le plus adapté pour classer la SNOMED CT (Dentler & Cornet, 2015).

Nous avons comparé les "mappings déclarés" (*i.e.*, les mappings créés de manière morphosyntaxique par les concepteurs de RxNorm entre les médicaments dans RxNorm et ceux dans la SNOMED CT) aux "mappings inférés" (mappings obtenus par la classification de la structure de la nouvelle RTO réalisée par le raisonneur).

Le tableau 1 décrit la distribution des concepts SNOMED CT en fonction de leur mappings aux concepts de RxNorm.

TABLE *1 – Distribution des concepts SNOMED CT décrivant le médicament, en fonction de leur mapping aux concepts RxNorm: comparaison des mappings inférés aux mappings définis par RxNorm (mappings déclarés)*

		Mappings déclarés		
		Présents	Absents	**Total**
Mappings inférés	Présents	1 892	85	1 977
	Absents	939	288	1 227
Total		2 831	373	3 204

L'intégration de RxNorm et de la SNOMED CT a permis de mettre en évidence l'intérêt d'exploiter les éléments définitionnels des concepts de chaque RTO. Certains mappings ont pu être retrouvés uniquement de manière inférée (85 mappings), témoignant de limites potentielles de la méthode morphosyntaxique utilisée par les concepteurs de RxNorm. En revanche, 939 mappings déclarés n'ont pas été retrouvés par le processus que nous avons mis en œuvre. Certaines différences étaient consécutives aux limites de notre processus d'intégration, telles que l'absence de conversion de certaines unités de mesure. Pour ce cas spécifique, une solution pourra être implémentée en utilisant l'UCUM [4] (Unified Code for Units of Measure). Nous avons par ailleurs identifié des incohérences dans les éléments définitionnels des concepts. En effet, il existe des différences de précision dans la définition de certains concepts. Par exemple, RxNorm n'utilise pas d'"unité de présentation" (unité comptable dans laquelle les médicaments sont présentés) pour décrire ses entités, contrairement à la SNOMED CT. Inversement, RxNorm utilise des "Qualitative Distinction" (*i.e.*, des étiquettes qui sont cliniquement pertinentes telles que "sans sucre") pour la description des médicaments, ce que ne fait pas la SNOMED CT. Des différences dans des éléments majeurs tels que le "Basis of strength" (substance de référence de la dose du médicament) ont également été trouvées, ce qui change fondamentalement la définition. La mise en évidence de ces différences a permis de proposer des pistes d'amélioration de la description du médicament dans les deux RTOs (Nikiema, 2018).

5 Cross-linking : cas des terminologies diagnostiques en cancérologie

Dans le domaine de la cancérologie, la réutilisation des données est confrontée à l'hétérogénéité des RTOs utilisées pour le codage des diagnostics. Afin de pallier cette difficulté, il est nécessaire de mettre en correspondance ces différentes RTOs et, en particulier, la CIM-10 (dixième révision de la Classification statistique Internationale des Maladies et des problèmes de santé connexes (WHO, 2011)) et la CIM-O3 (la troisième révision de la Classification Internationale des Maladies pour l'Oncologie (Fritz *et al.*, 2000)). Ces deux RTOs sont utilisées

4. https://unitsofmeasure.org/trac

de manière différente pour coder des diagnostics : la CIM-10 les décrit en tant que tels tandis que la CIM-O3 décrit des lésions histologiques et des localisations anatomiques, qui sont représentées suivant deux axes distincts et peuvent être combinées. Les notions représentées par ces RTOs étant distinctes, nous avons réalisé un processus de cross-linking entre elles en utilisant la SNOMED CT comme support.

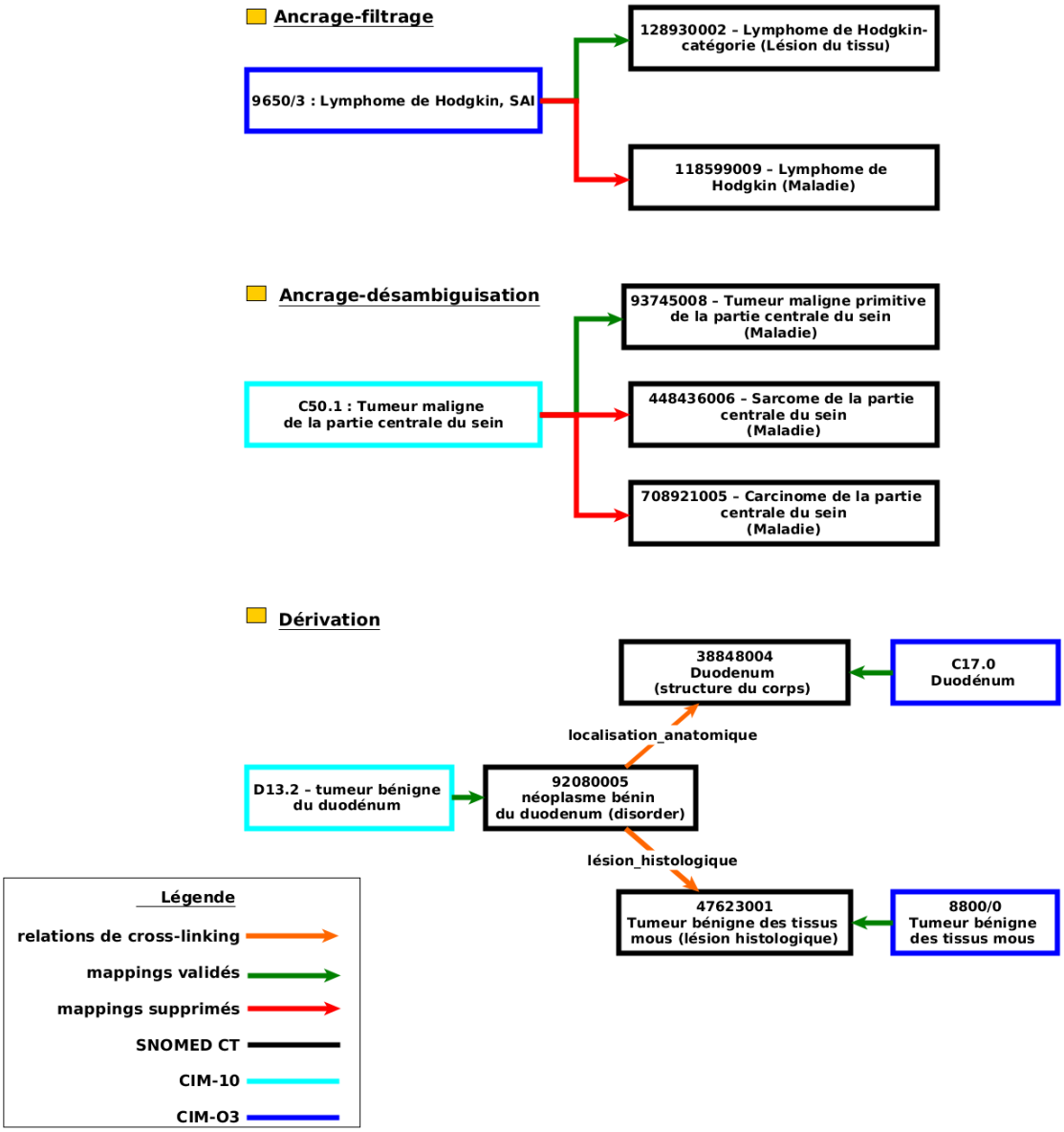


FIGURE 2 – *Exemple de cross-linking basé sur les relations transversales décrites dans la SNOMED CT*

Pour cela, deux étapes ont été mises en œuvre. Premièrement, la CIM-10 et la CIM-O3 ont été alignées à la SNOMED CT. Cette étape d'alignement, qualifiée d'**ancrage**, vise à rechercher des mappings d'équivalence entre les concepts de la CIM-10 et de la CIM-O3 et ceux de la SNOMED CT. La structure de la SNOMED CT a servi à : (1) filtrer les mappings incorrects (notamment les mappings entre des concepts de maladie et d'anatomie), et (2) désambiguïser les mappings multiples. La deuxième étape, dite de **dérivation**, a consisté

à établir des mappings complexes entre un concept CIM-10 et une paire de concepts CIM-O3. Tout d'abord, nous avons cherché dans la SNOMED CT les relations transversales pouvant lier les concepts de la CIM-10 et ceux de la CIM-O3. Nous avons ainsi identifié *finding_site* pour associer les diagnostics et les localisations anatomiques et *associated_morphology* entre les diagnostics et les lésions histologiques. Nous avons ensuite construit une structure inférée de la SNOMED CT grâce au raisonneur ELK (Dentler & Cornet, 2015). Sur la base des relations transversales identifiées, nous avons repéré les concepts CIM-10 équivalents à une combinaison de concepts CIM-O3 de morphologie et de topographie. Notons que l'identification d'inférences erronées dans l'étape de dérivation a permis de détecter des inconsistances dans la SNOMED CT.

Le cross-linking résulte en ce qui a été appelé dans la littérature des mappings complexes (Thieblin *et al.*, 2019). On parle de mapping complexe quand la relation de mapping est établie entre deux éléments dont au moins un des éléments n'est pas une simple entité. Ainsi, pour aller plus loin que l'établissement de mappings simples entre un code CIM-10 lié à un code CIM-O3 de topographie par une relation transversale, nous avons identifié des relations d'équivalence (lorsque le concept SNOMED CT était défini) ou de subsomption (lorsque le concept SNOMED CT était primitif) entre un concept CIM-10 et un couple de code CIM-O3. Nous avons automatiquement dérivé 86% (892/1032) des concepts CIM-O3 morphologiques avec 38% (127/330) de concepts CIM-O3 topographiques et 24% (203/852) des concepts CIM-10. La dérivation, analysée manuellement, a permis d'identifier des erreurs dans la hiérarchie de la SNOMED CT. Par exemple, elle a mis en évidence une relation de subsomption erronée entre les concepts 20955008-*insuline malin* et 3898006-*néoplasme bénin* (version de janvier 2017). Cette erreur a depuis été corrigée lors de la mise à jour de la SNOMED CT.

En conclusion, le cross-linking a permis de mettre en évidence l'intérêt d'utiliser une RTO de support pour les tâches suivantes : (i) la correction de mappings erronés lors de la phase d'ancrage, (ii) la découverte de mappings impliquant des relations transversales, et (iii) l'audit indirect de la RTO de support lorsque des inférences erronées ont été identifiées.

6 Intérêts de l'utilisation d'une RTO de support

Pour l'alignement et l'intégration, les stratégies appliquées dans cet article et dans la littérature se basent essentiellement sur le calcul de mesures de similarité entre entités provenant de différentes RTOs. Ces mesures de similarité sont généralement calculées d'après des techniques morphosyntaxiques, structurelles et sémantiques. Il est important de souligner que les similarités obtenues peuvent donner lieu à des interprétations erronées : ce sont les **conflits sémantiques** (Ngo *et al.*, 2013). Or, les conflits sémantiques ne sont pas tous résolus par les processus d'alignement et d'intégration.

Les méthodes morphosyntaxiques, consistant à retrouver des similarités entre les libellés des entités, sont souvent utilisées en premier lors de la création automatique de mappings (Aleksovski *et al.*, 2006; Shvaiko & Euzenat, 2013). Ces méthodes sont confrontées au risque de survenue de **conflits de nomenclature**, qui sont consécutifs aux similarités ou dissimilarités incorrectes entre des termes utilisés pour désigner les entités des RTOs. Ainsi :

— dans les cas d'homonymie, des mappings sont établis de manière erronée entre des concepts différents. Cette situation a été illustrée par les mappings qu'il était nécessaire de filtrer entre les concepts de la RTO locale et ceux de la LOINC.
— dans les cas de synonymie, certains concepts pourtant équivalents ne sont pas mappés.

Les méthodes structurelles sont habituellement utilisées après les méthodes morphosyntaxiques. Elles consistent à calculer le niveau de chevauchement des instances ou la proximité taxonomique des concepts présents dans les RTOs. Ces stratégies peuvent résoudre des cas de synonymie (Aleksovski *et al.*, 2006). Ainsi, à partir de la structure de la SNOMED CT et de RxNorm, des mappings ont pu être établis entre des concepts qui n'avaient pas été mappés par des méthodes morphosyntaxiques (Table 1). Cependant, les méthodes structurelles étant tributaires de la qualité de la structure des RTOs, elles sont sujettes aux conflits d'échelle et de confusion. Les **conflits d'échelle** apparaissent lorsqu'il y a une différence de granularité

dans les définitions des notions représentées (*e.g.*, absence d'"unités de présentation" dans la définition des concepts de RxNorm). Les **conflits de confusion** sont dus à des définitions contradictoires (*e.g.*, la signification différente de "Basis of strength" entre les concepts RxNorm et ceux de la SNOMED CT).

Les méthodes sémantiques décrites dans la littérature consistent à utiliser un support de connaissances. A partir des mappings avec une RTO de support, les entités de celles-ci servent à établir des ponts entre les RTOs à mettre en correspondance. Par exemple, à partir d'un concept de l'UMLS, il est possible de retrouver tous les concepts des RTOs intégrées dans l'UMLS qui sont censées décrire la même notion (Dhombres & Bodenreider, 2016; Mougin *et al.*, 2011; Kim *et al.*, 2012). Ainsi, les stratégies proposées dans la littérature permettent de réaliser l'alignement ou l'intégration de RTOs. Néanmoins, parce que ces stratégies se limitent à la recherche d'entités équivalentes ou reliées hiérarchiquement entre différentes RTOs, il n'est pas possible de les utiliser pour relier des entités décrivant des notions différentes mais complémentaires. Le processus permettant de prendre en compte ces limites est ce que nous avons qualifié de cross-linking. Celui-ci est basé sur une méthode combinant les méthodes de calcul de similarité pour répondre aux problématiques de conflits de confusion et d'échelle, tout en établissant des correspondances entre des entités différentes via des relations transversales (Nikiema *et al.*, 2017a; Jouhet *et al.*, 2017). Cette méthode repose sur l'utilisation d'une RTO de support et consiste en deux étapes : l'ancrage et la dérivation.

Comme les processus d'alignement et d'intégration que nous avons mis en place, l'ancrage à une RTO formelle permet la mise en place de procédures de validation des mappings basée sur la structure de ce support. La RTO de support doit disposer d'une structure formelle et être, au mieux, une ontologie pour une stratégie de mise en correspondance optimale. Les RTOs de support apportent des éléments définitionnels aux entités participant aux mappings, ce qui permet de s'affranchir de la qualité des structures des RTOs à relier. En effet, l'ancrage apporte des synonymes (Aleksovski *et al.*, 2006) et supprime des mappings erronés (Pesquita *et al.*, 2013) (comme illustré dans la figure 2 avec le filtrage en cas de conflit de confusion et la désambiguisation en cas de conflit d'échelle). Dans le cross-linking, la dérivation est l'étape essentielle qui permet d'améliorer l'organisation entre RTOs en reliant les entités différentes par des relations transversales grâce à la structure de la RTO de support.

7 Conclusion

Notre étude présente trois processus permettant d'utiliser conjointement des RTOs biomédicales hétérogènes. Deux aspects résument l'intérêt d'exploiter une RTO de support dans ce cadre : (1) la résolution des différents conflits sémantiques en procédant au filtrage de mappings erronés et à la désambiguïsation de mappings multiples, et (2) la possibilité d'établir automatiquement des mappings complexes entre des entités différentes mais décrivant des notions complémentaires.

Références

ABBURU S. (2012). A survey on ontology reasoners and comparison. *International Journal of Computer Applications*, **57**(17), 33–39.

AGENCY E. M. (2016). Introduction to ISO identification of medicinal products, spor programme.

ALEKSOVSKI Z., KLEIN M., TEN KATE W. & VAN HARMELEN F. (2006). Matching unstructured vocabularies using a background ontology. In S. STAAB & V. SVÁTEK, Eds., *Managing Knowledge in a World of Networks*, volume 4248, p. 182–197. Springer Berlin Heidelberg.

BODENREIDER O., CORNET R. & VREEMAN D. J. (2018). Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. *Yearbook of medical informatics*, **27**(01), 129–139.

BODENREIDER O. & JAMES J. (2018). The new SNOMED CT international medicinal product model. In *Proceedings of the 7h International Conference on Biological Ontology (ICBO 2018), Corvallis, Oregon, USA, August 7-10, 2018.*

BOURIGAULT D., AUSSENAC-GILLES N. & CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, **18**(1), 87–110.

DA SILVA C. F., MÉDINI L., GHAFOUR S. A., HOFFMANN P., GHODOUS P. & LIMA C. (2006). Semantic interoperability of heterogeneous semantic resources. *Electronic Notes in Theoretical Computer Science*, **150**(2), 71–85.

DENTLER K. & CORNET R. (2015). Intra-axiom redundancies in SNOMED CT. *Artificial Intelligence in Medicine*, **65**(1), 29–34.

DHOMBRES F. & BODENREIDER O. (2016). Interoperability between phenotypes in research and healthcare terminologies—investigating partial mappings between HPO and SNOMED CT. *Journal of Biomedical Semantics*, **7**(1), 3.

DIALLO G. (2014). An effective method of large scale ontology matching. *Journal of Biomedical Semantics*, **5**(1), 44.

EUZENAT J., MEILICKE C., STUCKENSCHMIDT H., SHVAIKO P. & TROJAHN C. (2011). Ontology alignment evaluation initiative: six years of experience. In *Journal on Data Semantics XV*, p. 158–192. Springer.

FRITZ A., PERCY C., SHANMUGARATNAM K., SOBIN L., PARKIN D. M. & WHELAN S. (2000). *International classification of diseases for oncology: ICD-O*. Geneva: World Health Organization, 3rd ed edition. OCLC: 248314653.

GARCIA I., KUSKA R. & SOMERMAN M. (2013). Expanding the foundation for personalized medicine: Implications and challenges for dentistry. *Journal of Dental Research*, **92**(7_suppl), S3–S10.

JOUBERT M., ABDOUNE H., MERABTI T., DARMONI S. & FIESCHI M. (2009). Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. In *AMIA Annual Symposium Proceedings*, p. 291–295.

JOUHET V., MOUGIN F., BRÉCHAT B. & THIESSARD F. (2017). Building a model for disease classification integration in oncology, an approach based on the National Cancer Institute thesaurus. *Journal of Biomedical Semantics*, **8**(1), 6.

KIM T. Y., COENEN A. & HARDIKER N. (2012). Semantic mappings and locality of nursing diagnostic concepts in UMLS. *Journal of Biomedical Informatics*, **45**(1), 93–100.

MARY M., SOUALMIA L. F. & GANSEL X. (2017). Formalisation de la terminologie loinc et évaluation de ses avantages pour la classification des tests de laboratoire. In *28es Journées francophones d'Ingénierie des Connaissances IC 2017*, p. 2–13.

MERABTI T., ABDOUNEB H., LECROQ T., JOUBERT M. & DARMONI S. J. (2009). Projection des relations SNOMED CT entre les termes de deux terminologies (CIM10 et SNOMED 3.5). *Risques, Technologies de l'Information pour les Pratiques Médicales*, p. 79–88.

MEYSTRE S. M., LOVIS C., BÜRKLE T., TOGNOLA G., BUDRIONIS A. & LEHMANN C. U. (2017). Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics*, **26**(01), 38–52.

MOUGIN F., DUPUCH M. & GRABAR N. (2011). Improving the mapping between MedDRA and SNOMED CT. In *Artificial Intelligence in Medicine*, p. 220–224. Springer.

NGO D., BELLAHSENE Z. & TODOROV K. (2013). Opening the black box of ontology matching. In *Extended Semantic Web Conference*, p. 16–30: Springer.

NIKIEMA J. N. (2018). Integrating RxNorm with medicinal products in SNOMED CT. Presentation at NLM/NIH, Bethesda. https://mor.nlm.nih.gov/pubs/alum/2018-nikiema-pres.pdf.

NIKIEMA J. N., JOUHET V. & MOUGIN F. (2017a). Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. *Journal of Biomedical Informatics*, **74**, 46–58.

NIKIEMA J. N., MOUGIN F. & JOUHET V. (2017b). Processus de prétraitement des libellés d'une terminologie d'interface. In *4e édition du Symposium sur l'Ingénierie de l'Information Médicale*, p. 95–103.

PESQUITA C., FARIA D., SANTOS E. & COUTO F. M. (2013). To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In *Proceedings of the 8th International Conference on Ontology Matching-Volume 1111*, p. 13–24: CEUR-WS. org.

PINTO H. S., GÓMEZ-PÉREZ A. & MARTINS J. P. (1999). Some issues on ontology integration. *In Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends*, **18**.

SHVAIKO P. & EUZENAT J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, **25**(1), 158–176.

STUDER R., BENJAMINS V. & FENSEL D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, **25**(1-2), 161–197.

THIEBLIN E., HAEMMERLÉ O., HERNANDEZ N. & TROJAHN C. (2019). Survey on complex ontology matching. *semantic web journal*.

WHO (2011). *International Statistical Classification of Diseases and Related Health Problems 10th revision*, volume 2. World Health Organization, 2010 edition.

Analyse de l'apprentissage humain dans la plateforme SIDES 3.0 : une approche basée sur la sémantique ⋆

Oscar Rodríguez Rocha, Catherine Faron Zucker

University Côte d'Azur, CNRS, Inria, I3S, France
oscar.rodriguez-rocha@inria.fr,faron@unice.fr

Résumé : SIDES 3.0 est un projet national français visant à fournir aux étudiants en médecine des services intelligents pour soutenir l'apprentissage en ligne dans le Système Intelligent d'Enseignement en Santé 3.0 (SIDES). La plateforme SIDES contient un grand nombre de ressources d'apprentissage annotées, notamment des questions de formation et d'évaluation, et recueille les traces d'apprentissage des étudiants. Ces annotations de ressources et traces d'apprentissage ont été intégrées sous la forme d'un graphe RDF, et enrichies grâce à des ontologies. Cet article présente les résultats de l'analyse de l'apprentissage des étudiants dans la plateforme SIDES en exploitant le graphe de connaissances associé en reposant sur les technologies du Web sémantique. Cette analyse est préliminaire à la conception et mise en oeuvre des fonctionnalités visant à permettre un apprentissage personnalisé et adaptatif sur la plateforme.

Mots-clés : e-Education, eHealth, Web sémantique, Représentation des connaissances et Raisonnement

1 Introduction

Depuis 2013, les facultés de médecine en France utilisent une plateforme nationale commune qui permet à leurs enseignants de créer et d'appliquer des tests d'évaluation locaux, qui sont ensuite partagés entre les universités afin de constituer une base de données nationale de tests de formation. Cette plateforme Web a été baptisée *SIDES* [1], acronyme de Système Informatisé Distribué d'Évaluation en Santé. Elle permet de préparer des étudiants en médecine aux Épreuves Classantes Nationales Informatisées (ECNi) depuis 2016.

Le projet français national *SIDES 3.0* vise à faire évoluer la plateforme SIDES vers une solution innovante baptisée Système Intelligent d'Enseignement en Santé 3.0 (SIDES 3.0), offrant des services intelligents centrés sur l'utilisateur, tels que : le suivi individuel, des tableaux de bord enrichis, des recommandations personnalisées, des corrections augmentées pour l'auto-évaluation, un environnement numérique normalisé de partage du savoir.

Pour atteindre cet objectif, le développement de *SIDES 3.0* s'appuie sur les modèles et technologies du Web sémantique et sur l'utilisation systématique des normes internationales en vigueur pour les métadonnées sur les ressources pédagogiques (MLR) [2] et les traces d'apprentissage (xAPI) [3], en les intégrant et les enrichissant par des ontologies.

Dans cet article nous présentons les résultats d'une analyse des ressources et de l'activité des étudiants de la plateforme SIDES, qui repose sur la conception d'un ensemble de requêtes SPARQL permettant d'interroger le graphe de connaissances construit à partir des données de la plateforme, en tenant compte de leur sémantique. Cette analyse axée sur les ressources et l'activité des étudiants dans la base de connaissances OntoSIDES, constitue la base pour concevoir et mettre en œuvre des fonctionnalités orientées à permettre un apprentissage personnalisé et adaptatif sur la plateforme *SIDES 3.0*. Après une brève présentation de travaux

⋆. Ce travail a été réalisé dans le cadre du projet DUNE SIDES 3.0 soutenu par l'Agence Nationale de Recherche (ANR-16-DUNE-0002-02).

1. http://side-sante.org
2. https://www.iso.org/standard/62845.html
3. https://xapi.com

connexes, nous commençons par une présentation du graphe de connaissances OntoSIDES, puis nous décrivons les résultats de l'analyse que nous avons menée de la plateforme SIDES, en termes de ressources pédagogiques et d'activité des apprenants, en montrant chaque fois quels types de requêtes permettent d'obtenir les résultats.

2 Travaux connexes

L'analyse de l'apprentissage est définie comme la mesure, la collecte, l'analyse et la présentation de données sur les élèves et leurs contextes, afin de comprendre et d'optimiser leur apprentissage (Ferguson, 2012). A notre connaissance, aucun des travaux existants ne présente une approche basée sur la sémantique pour l'analyse de l'apprentissage dans le domaine de la formation médicale. Des travaux connexes peuvent être trouvés dans l'analyse de l'apprentissage en exploitant les technologies du Web sémantique.

Dans (d'Aquin & Jay, 2013), les auteurs présentent une méthode qui exploite les connaissances externes disponibles sur le LOD pour faciliter l'interprétation des résultats d'exploration de données non sémantiques, en créant automatiquement une structure de navigation et d'exploration dans les résultats. Les résultats de l'exploration de données sont présentés de manière compatible avec une représentation LOD, puis sont liés aux sources du LOD existantes de sorte que l'analyste peut facilement explorer les résultats enrichis. Comparée à cette approche, celle que nous proposons n'exploite pas le LOD pour l'interprétation mais directement des données sémantiques dans le graphe OntoSIDES et pour cette même raison, notre analyse de l'apprentissage ne repose pas sur un algorithme d'exploration de données, mais sur des requêtes SPARQL.

MeLOD (Fulantelli *et al.*, 2013) est un environnement mobile conçu pour prendre en charge, via l'utilisation d'appareils mobiles, les expériences d'apprentissage informelles lors de la visite d'une ville. MeLOD exploite les technologies du Web sémantique pour prendre en charge les expériences d'apprentissage mobile et soutient l'analyse des apprentissages en fournissant des outils spécifiques pour analyser les activités des élèves. Comparée à ces travaux, notre approche s'intéresse à l'apprentissage dans le domaine de la médecine, ce qui change complètement l'objectif et le but de l'analyse d'apprentissage.

Dans (Softic *et al.*, 2013), les auteurs présentent les résultats d'une analyse des activités d'apprentissage basée sur le comportement des utilisateurs dans leur environnement d'apprentissage personnel à l'Université de technologie de Graz. Ils utilisent les technologies du Web sémantique pour mettre en place un tableau de bord d'analyse d'apprentissage pour la visualisation de métriques. La création d'un tableau de bord avec des métriques d'apprentissage pour les étudiants n'est pas l'objectif final de notre travail, mais c'est une perspective que nous discuterons avec les médecins impliqués dans le projet car cela pourrait aider les étudiants à améliorer leur expérience d'apprentissage.

Dans (Dietze *et al.*, 2017) les auteurs présentent le jeu de données LAK en RDF, un corpus de travaux de recherche dans les domaines de Learning Analytics et Educational Data Mining qui permet l'investigation et l'analyse de l'évolution des disciplines scientifiques et la validation de méthodes et d'outils scientométriques.

3 OntoSIDES

OntoSIDES (Palombi *et al.*, 2019) est un graphe de connaissances qui comprend une ontologie de domaine et un ensemble de déclarations factuelles sur des entités manipulées par la plateforme SIDES, reliant celles-ci aux classes et propriétés de l'ontologie. L'ontologie de domaine est représentée en OWL et les connaissances factuelles dans le modèle RDF. Il est ainsi possible d'interroger OntoSIDES avec le langage de requête standard SPARQL. Le graphe de connaissances OntoSIDES a été généré automatiquement à partir de la base de données relationnelles de la plateforme SIDES, et en enrichissant ces données à l'aide de l'ontologie développée. La version actuelle de l'ontologie OntoSIDES contient 52 classes et 50 propriétés. Les classes suivantes sont centrales dans la modélisation :

Action (`sides:action`) la classe des actions possibles des étudiants lorsqu'ils interagissent avec les ressources pédagogiques de la plateforme SIDES. Par exemple, avec la sous-classe `sides:action_to_answer` il est possible de caractériser l'action de sélectionner la proposition d'une réponse à une question.

Content (`sides:content`) la classe racine de la hiérarchie des types de ressources disponibles dans la plateforme SIDES. La classe des questions (`sides:question`), celle des propositions de réponse à une question (`sides:proposal_of_answer`) et celle des réponses (`sides:answer`) d'un étudiant à une question sont des sous-classes de `sides:content`. La Figure 2 présente un graphe RDF décrivant une action de réponse à une question d'un étudiant, qui utilise ces trois classes.

Referential entity (`sides:referential_entity`) la classe des éléments de référence du programme d'éducation français en médecine publié par le Ministre de l'Enseignement Supérieur.

Medical schools (`sides:institute`) la classe des universités et facultés de médecine dans lesquelles des plateformes locales SIDES sont déployées.

Person (`sides:person`) la classe des personnes impliquées dans les études de médecine. Ses sous-classes correspondent aux rôles spécifiques des utilisateurs de la plateforme SIDES : par exemple, la classe `sides:student` est une sous-classe de `sides:person`.

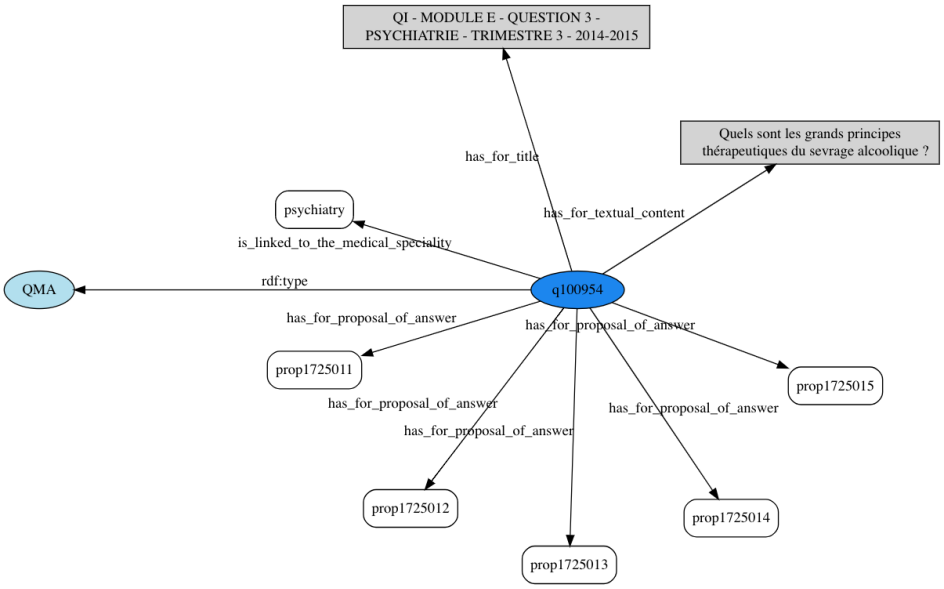


FIGURE *1 – Graphe RDF décrivant une question à réponse multiple (QMA)*

4 Analyse des ressources de la plateforme SIDES

Cette section vise à fournir une caractérisation du contexte de l'apprenant à travers des informations quantitatives et qualitatives sur les ressources présentes dans la plateforme SIDES. Ces informations ont toutes été calculées à l'aide de requêtes SPARQL appliquées sur le graphe OntoSIDES, dont certaines sont fournies dans la suite.

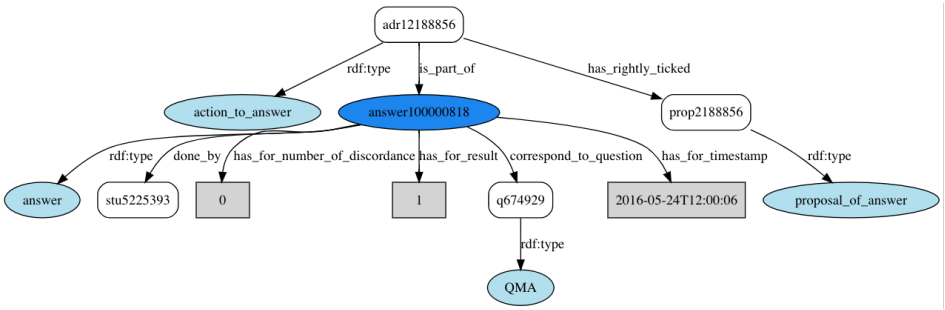


FIGURE *2 – Graphe RDF décrivant une réponse à une question*

4.1 Analyse des questions dans le graphe OntoSIDES

Le graphe de la plateforme SIDES contient actuellement un total de 590,654 questions différentes réparties en 4 catégories comme représenté sur la Figure 3. Cette distribution des questions a été calculée à l'aide de la requête 1. Comme cela apparaît clairement sur la Figure 3, la répartition des questions selon ces catégories n'est pas uniforme : Une grande majorité des questions dans la plateforme SIDES sont des questions à réponses multiples (QMA) : 467,498 questions, soit 79 ;1% ; il y a 81,155 questions à réponse unique (QUA), soit 13,7%, 40,249 questions rédactionnelles ouvertes courtes (QSOA), soit 6,8%, et seulement 1,752 questions de test de concordance de script (TCS), soit 0,3%.

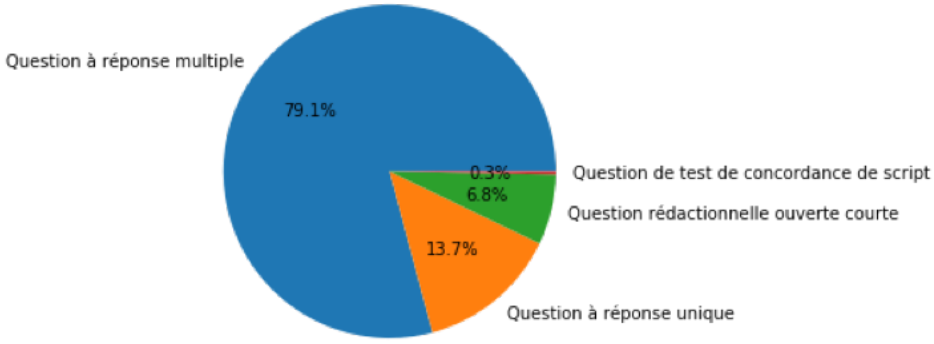


FIGURE *3 – Diagramme circulaire des catégories de questions*

```
SELECT ?type ?label (count(DISTINCT ?question) as ?count)
WHERE { ?question rdf:type ?type . ?type rdfs:subClassOf sides:question .
        ?type rdfs:label ?label FILTER (lang(?label) = 'fr')
} GROUP BY ?type ?label ORDER BY DESC(?count)
```

Query 1 – Requête SPARQL pour calculer la distribution des questions par type

Les descriptions de ces questions ne sont pas homogènes et souvent incomplètes. Notamment, les spécialités et objectifs d'apprentissage auxquelles les questions sont relatives qui permettent de bien caractériser le contexte d'apprentissage sont très peu mentionnés : 50,550 questions QMA, soit seulement 10.81%, sont liées à une spécialité médicale (i.e. décrites avec la propriété `sides:is_linked_to_speciality`), et 54,497 questions, soit seulement 11,66%, sont liées à un à un objectif d'apprentissage (i.e. décrites avec la propriété `sides:is_linked_to_ECN_referential_entity`). De façon similaire, 5,181 questions QUA, soit 6.38%, sont liées à une spécialité et 5,912 questions QUA, soit

7,28%, sont liées à un objectif d'apprentissage ; 1,034 questions QSOA, soit 2.57%, sont liées à une spécialité et 1,461 questions QSOA, soit 3,63% sont liées à un à un objectif d'apprentissage. Enfin, aucune question TCS n'est liée à une spécialité ou un objectif d'apprentissage.

4.2 Analyse des spécialités dans le graphe OntoSIDES

La plateforme SIDES contient actuellement des questions relatives à 31 spécialités médicales (instances de la classe `sides:speciality`). Le tableau 1 présente la répartition des questions par spécialité. Il a été construit à l'aide de la requête 2 qui calcule pour chaque spécialité le nombre de questions associées (une question peut être associée à plusieurs spécialités). Ce tableau montre notamment que la répartition des questions par spécialité n'est pas uniforme, variant de 189 pour la spécialité *Toxicologie* à 5440 pour la spécialité *Pédiatrie*, avec une moyenne de 2242 questions, et un écart type de 1283.

	QMA	QUA	QSOA	Total	%
Pédiatrie	4867	516	57	5440	7.83
Maladies infectieuses	4126	487	73	4686	6.74
Cardio-vasculaire	3550	300	86	3936	5.66
Endocrinologie - Métabolisme - Nutrition	3276	233	38	3547	5.10
Cancérologie - Radiothérapie	3042	300	75	3417	4.92
...					
Chirurgie maxillo-faciale	649	77	9	735	1.06
Neurochirurgie	553	51	13	617	0.89
Médecine du travail	496	59	8	563	0.81
Addictologie	286	43	9	338	0.49
Toxicologie	176	13	0	189	0.27

TABLE *1 – Nombre et proportion de questions associées à chaque spécialité*

```
SELECT ?speciality ?q_type (COUNT(DISTINCT ?question) AS ?questions)
WHERE { ?question rdf:type ?question_type .
        ?q_type rdfs:subClassOf sides:question .
        ?question sides:is_linked_to_the_medical_speciality ?speciality  .
} GROUP BY ?speciality ?q_type ORDER BY ASC(?speciality)
```

Query 2 – Requête SPARQL pour calculer le nombre de questions de chaque type associées à chaque spécialité

4.3 Analyse des objectifs d'apprentissage dans le graphe OntoSIDES

La plateforme SIDES contient actuellement des questions relatives à 362 objectifs d'apprentissage (instances de la classe `sides:ECN_learning_objective`) et 921 sous-objectifs d'apprentissage (instances de `sides:ECN_learning_sub_objective`). Ils ont été dénombrés avec la requête 3 qui prend en compte les relations de subsomption entre objectifs.

```
SELECT ?type ?label (count(DISTINCT ?lo) as ?count)
WHERE { ?lo rdf:type ?type . ?type rdfs:subClassOf* sides:ECN_referential_entity .
        ?type rdfs:label ?label FILTER (lang(?label) = 'fr')
} GROUP BY ?type ?label
```

Query 3 – Requête pour calculer le nb d'objectifs et sous objectifs d'apprentissage

Des requêtes similaires à celles utilisées pour l'analyse des spécialités permettent d'analyser la répartition des questions par objectifs d'apprentissage. De même que la répartition des questions par spécialité, celle des questions par objectif n'est pas uniforme, le nombre de questions associées à un objectif d'apprentissage variant de 7 pour l'objectif *Dopage* à 1651

pour l'objectif *Prescription et surveillance des classes de médicaments les plus courantes chez l'adulte et chez l'enfant*, avec une moyenne de 196 questions, et un écart type de 149.

5 Analyse de l'activité des étudiants dans le graphe OntoSIDES

Cette section fournit le résultat de l'analyse de l'activité des apprenants sur la plateforme SIDES réalisée en interrogeant le graphe OntoSIDES avec des requêtes SPARQL dédiées. Etant données les contraintes de longueur d'article imposée, nous nous sommes concentrés ici sur l'analyse de l'activité selon les spécialités. Un travail similaire a été conduit sur l'analyse de l'activité par objectif d'apprentissage.

5.1 Analyse de l'activité des étudiants par question

À ce jour, 64,957 étudiants (instances de la classe `sides:student`) sont identifiés sur la plateforme SIDES, mais seuls 41,442 étudiants on réalisé au moins une action, soit 63%. On constate que l'activité des étudiants actifs n'est pas uniforme, le nombre total de réponses données par étudiant variant de 1 à 62,015.

Le graphe OntoSIDES contient la description de 100,812,181 réponses à 456,854 questions, donc seules 77,34% des questions ont reçu au moins une réponse et chaque question a reçu en moyenne 221 réponses.

5.2 Analyse de l'activité des étudiants par spécialité

Toutes les spécialités ont été abordées par au moins un étudiant et en moyenne 17,600 étudiants ont abordé au moins une question de chaque spécialité. Cependant ce nombre n'est pas uniforme, variant de 6,111 pour *Toxicologie* à 24,461 pour *Maladies infectieuses*, avec un écart-type de 4,420.

Pour analyser plus finement l'activité des étudiants selon les spécialités, nous avons également calculé à l'aide de la requête 4 le nombre de réponses à des questions par spécialité et le nombre moyen de réponses à des questions par étudiant dans chaque spécialité. La figure 4 montre les résultats obtenus à partir de cette requête, les spécialités étant triées par ordre décroissant du nombre moyen de réponses par étudiant. On constate ainsi que *Maladies infectieuses* est la spécialité ayant reçu le plus grand nombre de réponses à des questions, mais que *Pédiatrie* est la spécialité avec le plus grand nombre de questions par étudiants. Une interprétation possible est que les étudiants intéressés par la *Pédiatrie* ont été les plus actifs dans cette spécialité. De même, les étudiants les moins actifs dans uns spécialité ont été ceux qui s'intéressaient à la *Toxicologie*.

```
SELECT ?label (COUNT(DISTINCT ?answer) AS ?answers) (COUNT(DISTINCT ?student) AS ?students)
WHERE { ?answer rdf:type sides:answer .
        ?answer sides:correspond_to_question ?question .
        ?answer sides:done_by ?student .
        ?question sides:is_linked_to_the_medical_speciality ?speciality  .
        { SELECT ?speciality (MIN(?duplicated_label) AS ?label)
          WHERE { ?speciality a sides:speciality .
                  ?speciality rdfs:label ?duplicated_label .
                  FILTER (lang(?duplicated_label) = "fr")
          } GROUP BY ?speciality ORDER BY ?label }
} GROUP BY ?label ORDER BY ?label
```

Query 4 – Requête pour calculer le nb d'étudiants et le nb de réponses par spécialité

Pour approfondir encore notre analyse, nous nous sommes également intéressés à la qualité des réponses des étudiants aux questions par spécialité. Nous nous sommes limités à l'analyse des réponses aux questions à réponse unique (QUA) dont le résultat est binaire (correct ou faux). Nous pourrions étendre l'analyse aux questions à réponses multiples (QMA), en considérant un seuil au-delà duquel considérer que le nombre d'options correctement sélectionnées pour une répondre à une question constitue une réponse correcte. La requête 5 permet de compter les nombres de réponses correctes et incorrectes à des questions QUA et la Figure 5 présente les résultats obtenus, les spécialités étant triées par ordre décroissant du

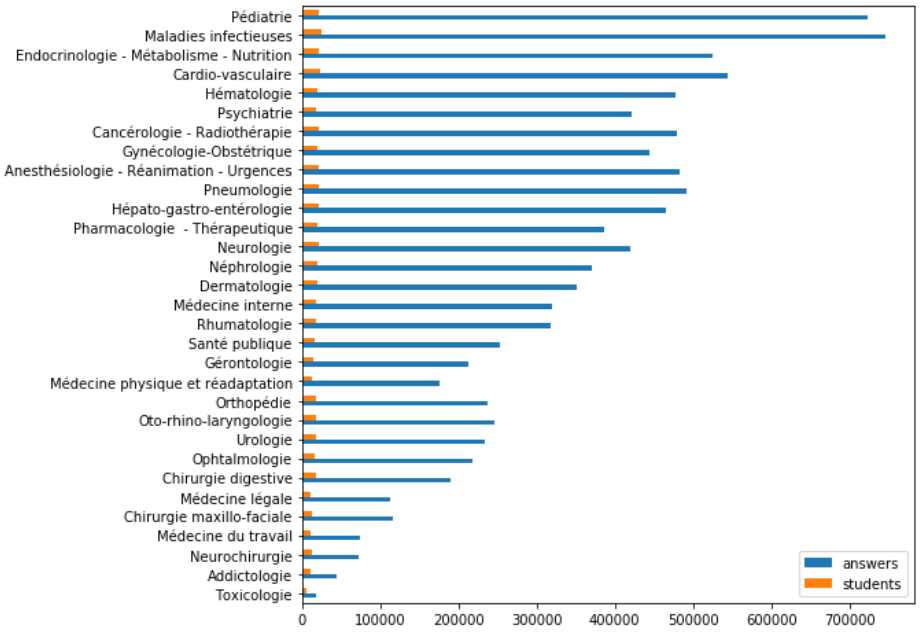


FIGURE *4 – Nombre de réponses et d'étudiants ayant répondu pour chaque spécialité. Les spécialités sont triées par ordre décroissant d'activité des étudiants mesuré par le nombre moyen de réponses par étudiant*

ratio entre le nombre de réponses correctes et le nombre de réponses incorrectes. On constate ainsi notamment que la spécialité *Maladies infectieuses* est celle ayant le plus grand nombre de réponses correctes, la spécialité *Médecine physique et réadaptation* est celle avec la plus grande proportion de réponses correctes, et la spécialité *Toxicologie* est celle avec la plus faible proportion de réponses correctes (il y a même davantage de réponses incorrectes que correctes). Ceci, combiné aux résultats de la requête précédente, montre que c'est la spécialité dans laquelle les étudiants répondent le moins et le moins bien. Une interprétation possible est que la spécialité *Toxicologie* est celle pour laquelle les concepts à acquérir sont les plus difficiles, tandis que la *Médecine physique et réadaptation* celle manipulant les concepts les plus simples. Une autre interprétation peut être donnée en terme de qualité des questions conçues pour chaque spécialité, les questions les moins bien répondues pouvant nécessiter une révision.

```
SELECT ?label (sum(if(?result = 1, 1, 0)) as ?corrects)
(sum(if(?result = 1, 0, 1)) as ?wrongs) (count(?answer) as ?answers)
WHERE { ?answer rdf:type sides:answer .
        ?answer sides:correspond_to_question ?question .
        ?question a sides:QUA .
        ?question sides:is_linked_to_the_medical_speciality ?speciality .
        ?answer sides:has_for_result ?result .
        { SELECT ?speciality (MIN(?duplicated_label) AS ?label)
          WHERE { ?speciality a sides:speciality .
                  ?speciality rdfs:label ?duplicated_label .
                  FILTER (lang(?duplicated_label) = "fr")
          } GROUP BY ?speciality ORDER BY ?label }
} ORDER BY DESC(?corrects)
```

Query 5 – Requête pour extraire le résultat des réponses aux questions pour chaque spécialité

6 Conclusions et perspectives

Nous avons présenté un premier travail d'analyse des ressources de la plateforme d'apprentissage SIDES et de l'activité des étudiants sur la plateforme, basé intégralement sur

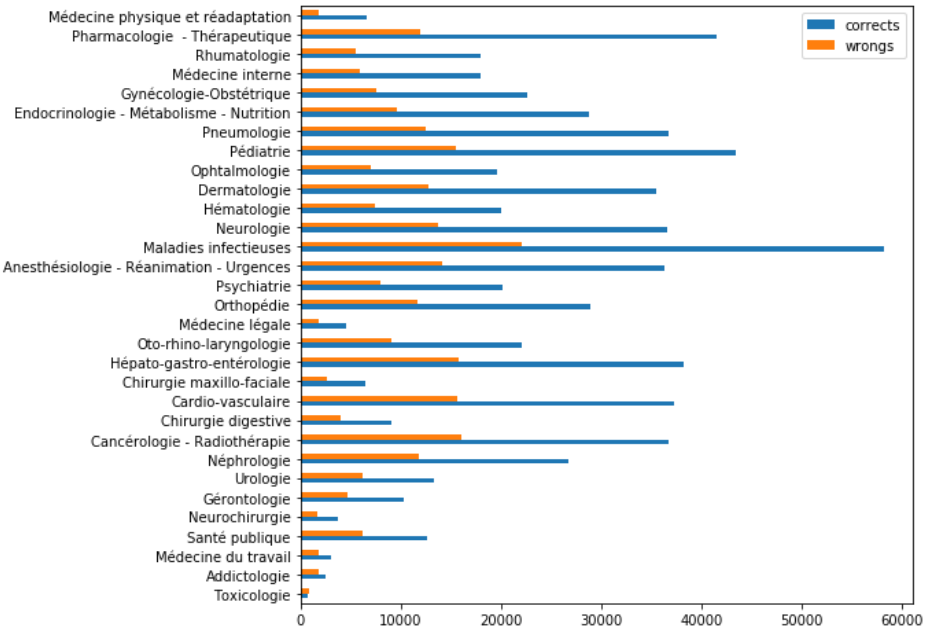


FIGURE *5 – Nombres de réponses correctes et incorrectes à des questions QUA pour chaque spécialité. Les spécialités sont triées par ordre décroissant de difficulté mesurée par le ratio entre les nombres de réponses correctes et incorrectes*

l'interrogation du graphe de connaissance OntoSIDES en RDF avec des requêtes SPARQL. Le résultat des analyses peut donner lieu à différentes interprétations possibles, en terme de suivi individuel des apprenants mais aussi en terme de retours sur la popularité, la difficulté, voire la qualité des ressources disponibles sur la plateforme. A court terme nous allons discuter et valider ces résultats et interprétations avec les médecins impliqués dans le projet. Les résultats de ce travail nous serviront pour concevoir et mettre en œuvre des fonctionnalités orientées vers un apprentissage adaptatif et personnalisé dans la plateforme, telles que la recommandation de questions en fonction du niveau de connaissances, des objectifs d'apprentissage et des spécialités médicales de l'étudiant.

Références

D'AQUIN M. & JAY N. (2013). Interpreting data mining results with linked data for learning analytics : Motivation, case study and directions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, p. 155–164, New York, NY, USA : ACM.

DIETZE S., TAIBI D. & D'AQUIN M. (2017). Facilitating scientometrics in learning analytics and educational data mining - the lak dataset. *Semantic Web*, **8**, 395–403.

FERGUSON R. (2012). Learning analytics : drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, **4**(5/6), 304–317.

FULANTELLI G., TAIBI D. & ARRIGO M. (2013). A semantic approach to mobile learning analytics. In *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, TEEM '13, p. 287–292, New York, NY, USA : ACM.

PALOMBI O., JOUANOT F., NZIENGAM N., OMIDVAR-TEHRANI B., ROUSSET M.-C. & SANCHEZ A. (2019). Ontosides : Ontology-based student progress monitoring on the national evaluation system of french medical schools. *Artificial Intelligence in Medicine*, **96**, 59 – 67.

SOFTIC S., TARAGHI B., EBNER M., DE VOCHT L., MANNENS E. & VAN DE WALLE R. (2013). Monitoring learning activities in ple using semantic modelling of learner behaviour. In *Human Factors in Computing and Informatics*, p. 74–90 : Springer Berlin Heidelberg.

PEPS, une plateforme de prévention cardiovasculaire orientée citoyen

Adrien Ugon[1,2], Hector Falcoff[3,4], Emmanuel Jobez[3,4], Madeleine Favre[3,4], Rosy Tsopra[5], Pierre Meneton[5], Marie-Christine Jaulent[5]

[1] ESIEE-PARIS, Noisy-le-Grand, France
adrien.ugon@esiee.fr

[2] LIP6, Sorbonne Université, CNRS UMR 7606, Paris, France
adrien.ugon@lip6.fr

[3] SOCIÉTÉ DE FORMATION THÉRAPEUTIQUE DU GÉNÉRALISTE, Paris, France
hector.falcoff@sfr.fr, e.jobez@yahoo.fr, docteurmfavre@gmail.com

[4] COLLÈGE DE LA MÉDECINE GÉNÉRALE, Paris, France

[5] LIMICS, INSERM UMRS 1142, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny, France UPMC Université Paris 6, Sorbonne Universités, Paris
rosytsopra@gmail.com, pierre.meneton@crc.jussieu.fr, marie-christine.jaulent@inserm.fr

Résumé : La plateforme PEPS a pour objectif de donner au citoyen des outils d'auto-évaluation du risque cardiovasculaire et la conception d'un plan personnalisé de prévention cardiovasculaire. Le risque global est analysé par quinze facteurs de risque, évalués par des questionnaires dédiés. Lorsque le facteur de risque a été identifié comme présent, des actions sont proposées pour réduire ce facteur de risque et le risque cardiovasculaire global. Co-construit par le patient et son médecin traitant, le plan de prévention personnalisé est un outil qui accompagne le citoyen dans des actions de prévention, en respectant ses motivations et ses préférences , respectant les principes du *patient empowerment*.
Une première évaluation ergonomique a permis de mettre en évidence que le nombre important de question pouvait être un frein à l'approche.

Mots-clés : Risque cardiovasculaire, Prévention, Patient empowerment, Moteur de règles.

1 Introduction

Les maladies cardiovasculaires représentent la première cause de décès en Europe (Townsend *et al.*, 2016). Des actions de prévention impliquant le citoyen permettent de réduire le risque cardiovasculaire (Kathuria-Prakash *et al.*, 2019) en s'inscrivant dans des démarches de *patient empowerment* dont les bénéfices pour le citoyen ont été démontrés (Kambhampati *et al.*, 2016).

Des approches globales permettent d'évaluer le risque cardiovasculaire dans son intégralité à partir d'un score. L'échelle de risque de Framingham-D'Agostino permet d'évaluer la morbidité et la mortalité cardiovasculaire ; il a été validé sur la population des États-Unis (D'Agostino *et al.*, 2008). En Europe, la société européenne de cardiologie (*European Society of Cardiology*) a proposé un score appelé SCORE (*Systematic Coronary Risk Estimation*) afin d'estimer le risque d'accident cardiovasculaire fatal à dix ans, chez des sujets âgés de 40 à 65 ans, non diabétiques et dont la tension artérielle ne dépasse pas 160/110.
Des études comparatives ont été menées entre l'échelle de Framingham-D'Agostino et SCORE et ont montré des différences significatives (Gómez-Marcos *et al.*, 2009), mettant en évidence la difficulté d'une approche globale.

Des travaux récents ont montré l'importance de facteurs de risque rarement pris en compte, comme le chômage, les troubles du sommeil ou la dépression dans le risque cardiovasculaire global (Meneton *et al.*, 2014), (Meneton *et al.*, 2016). Ces facteurs de risque

ne sont pas considérés dans les approches globales citées précédemment.

Les pathologies cardiovasculaires sont associées à de nombreux facteurs de risque, sur lesquels il est possible d'agir séparément dans des objectifs de prévention. En France, la Haute Autorité de Santé (HAS) publie des guides de bonne pratique cliniques dédiés aux facteurs de risque médicaux comportant des sections dédiées à la prévention (Hau, 2014b), (Hau, 2017a), (Hau, 2017b), (Hau, 2016), (Hau, 2012), (Hau, 2012), (Hau, 2014a). Ces guides ne couvrent pas l'intégralité des facteurs de risque cardiovasculaires reconnus.

L'objectif de ce projet est de concevoir une plateforme orientée citoyen pour lui permettre de co-construire, soit tout seul, soit en coopération avec son médecin traitant, un plan de prévention cardiovasculaire. La structure de cette plateforme a été présentée dans des travaux précédents (Ugon *et al.*, 2018b). Après identification d'un profil cardiovasculaire, permettant de déterminer l'absence ou la présence de différents facteurs de risque, le citoyen se verra proposer des boîtes à outils avec des actions à mettre en place et des recommandations destinées à réduire son risque cardiovasculaire global.

L'objectif de cet article est de présenter la plateforme PEPS, une plateforme orientée pour le citoyen pour lui permettre de construire seul, ou co-construire avec son médecin traitant, un plan personnalisé de prévention cardiovasculaire. La section 2 présente la méthode en quatre parties ; la section 3 expose les résultats préliminaires à ce projet suivie d'une discussion dans la section 4. Finalement, une conclusion est proposée dans la section 5.

2 Méthodes

2.1 Détermination des facteurs de risque

Un groupe de travail a déterminé une liste de 15 facteurs de risque cardiovasculaire, regroupés en quatre catégories.

La première catégorie rassemble cinq facteurs de risque liés au mode de vie et au comportement : la consommation de tabac, la consommation d'alcool, l'activité physique insuffisante, l'alimentation déséquilibrée et les troubles du sommeil. La deuxième catégorie comporte quatre facteurs de risque psycho-socio-environnementaux : le stress, la dépression, la position socio-professionnelle défavorable et la pollution atmosphérique. La troisième catégorie concerne quatre facteurs de risque médicaux classiques : l'obésité, l'hypertension artérielle, le diabète et les dyslipidémies. La dernière catégorie concerne deux pathologies augmentant le risque cardiovasculaire : l'insuffisance rénale chronique et la maladie inflammatoire.

2.2 Construction d'un moteur de règles

Des travaux précédents sur la formalisation de ces guides de bonne pratique clinique de la HAS en des arbres décisionnels a permis d'obtenir un système d'aide à la décision retournant l'existence, ou l'absence, d'un facteur de risque chez un citoyen, à partir de ses données démographiques et cliniques (Ugon *et al.*, 2018a). Ces arbres de décision ont ensuite été implémentés dans un système à base de règles sous la forme d'une API installable localement.

Le moteur de règles permet d'apporter une conclusion sur l'évaluation d'un facteur de risque. Quatre valeurs sont utilisées : « présent » si le facteur de risque est présent, « absent » si le facteur de risque est absent, « non évalué » si le citoyen a fait le choix de ne pas évaluer sa situation selon ce facteur de risque et « non évaluable », si les informations saisies par le citoyen sont insuffisantes pour permettre de conclure sur ce facteur de risque.

2.3 Élaboration de questionnaires

Afin de pouvoir alimenter ce moteur de règles, un questionnaire a été rédigé par des experts médicaux, avec l'assistance d'experts en informatique médicale. Son but est de collecter l'ensemble des données nécessaires à la prise de décision par le moteur de règles. Il compte 108 questions, certaines de ces questions étant le score obtenu dans d'autres questionnaires, tels que le questionnaire CAST, évaluant la consommation de cannabis, ou le score EPICES évaluant la précarité. Chacune des réponses a été codée afin de pouvoir être interprétée sans ambiguïté et être interopérable avec les systèmes d'information des différents logiciels métiers auxquels cet outil se propose d'être adjoint.

Afin de pouvoir identifier les autres facteurs de risque, d'autres questionnaires ont été identifiés, avec l'aide d'un épidémiologiste et de spécialistes médicaux des différents domaines : le stress est évalué à l'aide de l'échelle de stress perçu de Cohen à 4 items (Cohen *et al.*, 1983) ; la dépression est évaluée avec le *Patient Health Questionnaire-9* (Kroenke *et al.*, 2001), comportant neuf questions ; l'apnée du sommeil est évaluée avec le questionnaire STOP-BANG (Chung *et al.*, 2016), évaluant huit critères , trois questions supplémentaires sont ajoutées pour identifier les autres troubles du sommeil ; l'activité physique insuffisante a été évaluée avec le questionnaire d'auto-évaluation du niveau d'activité physique hebdomadaire de J. Ricci et L. Gagnon, de l'université de Montréal, modifié par F. Laureyns et JM. Séné (Ricci & Gagnon, 2009) comportant neuf questions ; la consommation de tabac est évaluée avec le test mini-Fagerström (Fagerström, 1978) comportant six questions ; l'alcool est évaluée avec un questionnaire à trois questions inspirées des recommandations officielles françaises sur la consommation de boissons alcoolisées (Service, 2017) ; l'obésité est évaluée à partir de l'Indice de Masse Corporelle (IMC), en posant deux questions au citoyen, sa taille et son poids ; l'exposition à la pollution est évaluée avec une seule question ; la position socio-professionnelle défavorable est évaluée par six questions ; finalement, l'alimentation déséquilibrée est évaluée avec un questionnaire élaboré par le groupe de travail en coopération avec une équipe experte en épidémiologie de la nutrition, et les recommandations du Programme National Nutrition Santé (PNNS).

Chacun de ces questionnaires standards est évalué à partir d'un score calculé à partir des réponses données. Des catégories sont associées par intervalle de score permettant de déterminer différents niveaux de gravité.

2.4 Conception d'un plan personnel de prévention

Après avoir répondu à différents questionnaires, le citoyen obtient des résultats personnalisés, l'information de son profil cardiovasculaire. Il sait, pour chacun des facteurs de risque pour lesquels il a répondu aux questions, si ce facteur est présent ou non.

Il peut ainsi choisir sur l'interface un — ou plusieurs — facteur de risque identifiés comme présent pour pouvoir agir. Un web service interroge alors une base de connaissance afin d'obtenir des actions à mettre en place pour réduire ce facteur de risque, et, en conséquence, réduire le risque cardiovasculaire global. Les actions sont regroupées afin d'être présentées sur deux niveaux hiérarchiques ; par exemple, les actions favorisant la lutte contre la sédentarité sont regroupées en deux catégories : « être plus actif au quotidien » et « pratiquer un activité sportive ». Chaque action est en plus étiquetée par une des trois catégories suivantes : « Faire », « Se documenter », « Consulter un professionnel ». De plus, une fréquence d'autoévaluation est associée, afin de savoir la régularité à laquelle cette action doit être réalisée, et donc évaluée, dans la réalisation de son plan de prévention par le citoyen. Certaines actions ne sont présentées que lorsque le citoyen est accompagné par son médecin traitant.

3 Résultats

Des maquettes de la plateforme ont été réalisées, intégrant les contenus scientifiques spécifiés par les groupes d'experts. Un aperçu des questionnaires est donné sur la figure 1.

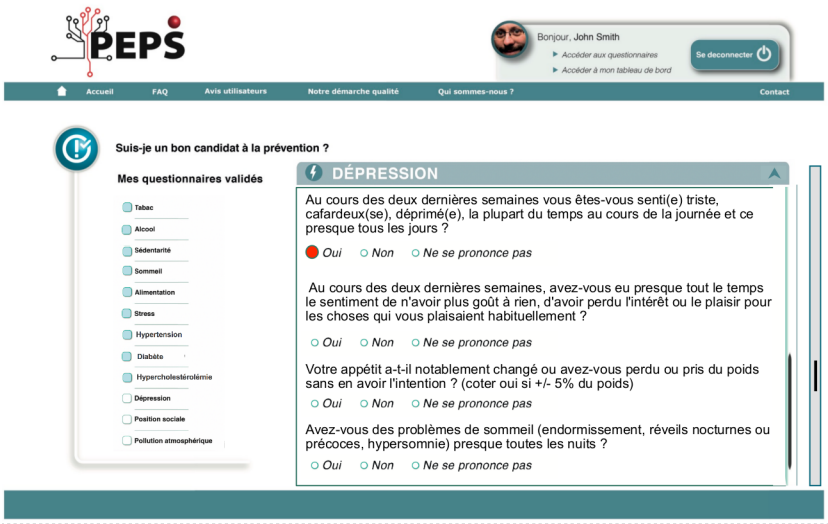


FIGURE *1 – Exemple de maquette de la plateforme PEPS*

Des évaluations préliminaires ont été conduites par des experts en évaluation ergonomique afin d'évaluer la faisabilité et l'ergonomie de l'interface de saisie. Des maquettes ont été présentées à des évaluateurs ; leur avis a été pris en note tout au long de l'évaluation.
Un groupe d'évaluateurs incluant des médecins généralistes a rempli les questionnaires correspondant à deux situations cliniques factices. Cette saisie a duré près de deux heures pour chacune des deux situations. Des difficultés ont été rencontrées pour répondre à certaines questions, l'information étant jugée complexe à trouver dans un logiciel métier et parfois ambiguë.

Les évaluateurs ont estimé que le nombre important de questions à répondre dans les différents questionnaires était un frein important à l'utilisation de l'outil. De plus, l'information à renseigner est parfois compliquée à obtenir, et son lien non évident avec les maladies cardiovasculaires ne motivent pas y répondre.

4 Discussion

L'approche de la plateforme PEPS met en évidence la difficulté d'une approche globale. L'adhésion du citoyen à une telle plateforme nécessite d'évaluer son risque cardiovasculaire par un nombre réduit de questions. L'identification et la séparation en différents facteurs de risque permet de se concentrer sur certains d'entre eux seulement, et de respecter les éléments de motivation du citoyen.
Certaines informations sont difficiles à obtenir par le citoyen seul, ce qui peut être une raison d'abandon.

La co-construction du plan de prévention entre le citoyen et son médecin traitant constitue une démarche innovante également du point de vue du médecin, qui va s'intéresser à des facteurs de risque habituellement non abordés en consultation. Les informations et actions recommandées de la base de connaissance offrent une garantie de rester dans le respect de la pratique de la *médecine basée sur les preuves*.

5 Conclusion

La plateforme PEPS est une plateforme orientée citoyen dédiée à la conception seul, ou la co-construction entre le citoyen et son médecin traitant d'un plan personnalisé de prévention cardiovasculaire. Le profil du citoyen est composé de quinze facteurs de risques, chacun étant évalué à « présent », « absent », « non évalué » ou « non évaluable ». L'identification des facteurs de risque se fait à l'aide de questionnaires remplis par le citoyen. Une partie des questions est ensuite injectée dans un moteur de règles, l'autre partie est soumise à un calcul de scores, l'ensemble permettant d'établir le profil cardiovasculaire.
En fonction de ce profil, le citoyen peut alors faire le choix d'un facteur de risque cariovasculaire pour lequel il éprouve une volonté d'agir. Des actions sont alors proposées, issues d'une base d'actions. Les évaluations ont mis en évidence la difficulté de remplir un très grand nombre de questions, obligeant à s'orienter vers une évaluation la plus succincte possible, pouvant éventuellement être complétée par le médecin traitant, si le citoyen le souhaite.
Des évaluations plus approfondies doivent être bientôt conduites.

Remerciements

Ce travail a été financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet PEPS DS0412–2016 (ANR-16-CE19-0018)

Références

(2012). *Arrêt de la consommation de tabac : du dépistage individuel au maintien de l'abstinence en premier recours*. Haute Autorité de Santé, 2 avenue du Stade de France - 93218 Saint-Denis La Plaine CEDEX.

(2014a). *Arrêt de la consommation de tabac : du dépistage individuel au maintien de l'abstinence en premier recours*. Haute Autorité de Santé, 2 avenue du Stade de France - 93218 Saint-Denis La Plaine CEDEX.

(2014b). *Prévention et dépistage du diabète de type 2 et des maladies liées au diabète*. Haute Autorité de Santé, 2 avenue du Stade de France - 93218 Saint-Denis La Plaine CEDEX.

(2016). *Prise en charge de l'hypertension artérielle de l'adulte*. Haute Autorité de Santé, 2 avenue du Stade de France - 93218 Saint-Denis La Plaine CEDEX.

(2017a). *Principales dyslipidémies : stratégies de prise en charge*. Haute Autorité de Santé, 2 avenue du Stade de France - 93218 Saint-Denis La Plaine CEDEX.

(2017b). *Évaluation du risque cardio-vasculaire*. Haute Autorité de Santé, 2 avenue du Stade de France - 93218 Saint-Denis La Plaine CEDEX.

CHUNG F., ABDULLAH H. R. & LIAO P. (2016). STOP-bang questionnaire. *Chest*, **149**(3), 631–638.

COHEN S., KAMARCK T. & MERMELSTEIN R. (1983). A global measure of perceived stress. *J Health Soc Behav*, **24**(4), 385–396.

D'AGOSTINO R. B., VASAN R. S., PENCINA M. J., WOLF P. A., COBAIN M., MASSARO J. M. & KANNEL W. B. (2008). General cardiovascular risk profile for use in primary care. *Circulation*, **117**(6), 743–753.

FAGERSTRÖM K.-O. (1978). Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. *Addictive Behaviors*, **3**(3-4), 235–241.

GÓMEZ-MARCOS M. A., MARTÍNEZ-SALGADO C., MARTIN-CANTERA C., RECIO-RODRÍGUEZ J. I., CASTAÑO-SÁNCHEZ Y., GINÉ-GARRIGA M., RODRIGUEZ-SANCHEZ E. & GARCÍA-ORTIZ L. (2009). Therapeutic implications of selecting the SCORE (european) versus the dAGOSTINO (american) risk charts for cardiovascular risk assessment in hypertensive patients. *BMC Cardiovascular Disorders*, **9**(1).

KAMBHAMPATI S., ASHVETIYA T., STONE N. J., BLUMENTHAL R. S. & MARTIN S. S. (2016). Shared decision-making and patient empowerment in preventive cardiology. *Current Cardiology Reports*, **18**(5).

KATHURIA-PRAKASH N., MOSER D., ALSHURAFA N., WATSON K. & EASTWOOD J. (2019). Young african american women's participation in an m-health study in cardiovascular risk reduction : Feasibility, benefits, and barriers. *European Journal of Cardiovascular Nursing*, p. 147451511985000.

KROENKE K., SPITZER R. L. & WILLIAMS J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, **16**(9), 606–613.

MENETON P., KESSE-GUYOT E., MÉJEAN C., FEZEU L., GALAN P., HERCBERG S. & MÉNARD J. (2014). Unemployment is associated with high cardiovascular event rate and increased all-cause mortality in middle-aged socially privileged individuals. *International Archives of Occupational and Environmental Health*, **88**(6), 707–716.

MENETON P., LEMOGNE C., HERQUELOT E., BONENFANT S., LARSON M. G., VASAN R. S., MÉNARD J., GOLDBERG M. & ZINS M. (2016). A global view of the relationships between the main behavioural and clinical cardiovascular risk factors in the GAZEL prospective cohort. *PLOS ONE*, **11**(9), e0162386.

RICCI J. & GAGNON L. (2009). Évaluation du niveau d'activité physique et de condition physique.

SERVICE A. I. (2017). Alcool : pour une consommation à moindre risque. `http://www.alcool-info-service.fr/alcool/consequences-alcool/consommation-a-risque`. Accessed : 2019-05-16.

TOWNSEND N., WILSON L., BHATNAGAR P., WICKRAMASINGHE K., RAYNER M. & NICHOLS M. (2016). Cardiovascular disease in europe : epidemiological update 2016. *European Heart Journal*, **37**(42), 3232–3245.

UGON A., HADJ BOUZID A. I., JAULENT M.-C., FAVRE M., DUCLOS C., JOBEZ E., FALCOFF H., LAMY J.-B. & TSOPRA R. (2018a). Building a knowledge-based tool for auto-assessing the cardiovascular risk. *Studies in Health Technology and Informatics*, **247**(Building Continents of Knowledge in Oceans of Data : The Future of Co-Created eHealth), 735–739.

UGON A., JOBEZ E., FALCOFF H., JAULENT M.-C., MENETON P., FAVRE M. & TSOPRA R. (2018b). Modular knowledge-based decision support system dedicated to a cooperative decision to prevent cardiovascular diseases. *Studies in Health Technology and Informatics*, **255**(Decision Support Systems and Education), 200–204.

Digital Implantable Gastric Stethoscope for the detection of early signs of acute cardiac decompensation in patients with chronic heart failure.

Cindy Dopierala[a, b], Pierre-Yves Guméry[a], Mohamed-Ridha Frikha [a], Jean-Jacques Thiébault [a], Philippe Cinquin [a,b], François Boucher[a]

[a]Univ. Grenoble Alpes, CNRS, Grenoble INP, VetAgro Sup, CHU Grenoble Alpes, TIMC-IMAG, F-38000 Grenoble, France*
**Campus Vétérinaire, BP 83,69280, Marcy 1'Etoile. France*
[b]SentinHealth, 38700 La Tronche, France

Abstract

To date, early warning signs of acute decompensated heart failure (ADHF) are not detected; medical treatment is adjusted only at the onset of symptoms leading to hospitalization. We develop an implantable and integrated device able to track various clinical and subclinical parameters with the aim of providing early detection of ADHF. Our originality lies in the choice of a gastric implantation site which has never been used before and opens a promising exploratory field associated with new potentially relevant parameters in the follow-up of HF. An in-vivo experimentation was run to validate its feasibility in two healthy pigs with a first prototype. Electrophysiological and mechanical signals were recorded over a two-week period, to conduct a first analysis of relevant parameters for ADHF detection such as heart rate variability and heart sounds. Promising preliminary results confirm the interest of considering the stomach as a strategic implantation site for cardio-respiratory monitoring.

Keywords:

Multimodal signal processing - Heart Failure – Heart Sounds – Electrocardiography

Introduction

Heart failure (HF) is a major public health concern, currently increasing because of the ageing populations and already affecting approximately 15 million people in Europe. Heart failure has also an important economic weight. Indeed, the expenses related to this syndrome are estimated at 2.9 billion euros in France in 2015 (2% of health expenditure), of which 75% is attributable to hospitalizations [1]. HF is a complex and heterogeneous syndrome whose mortality in the medium time is high. The Framingham study has followed patients with chronic HF during fifteen years (median), the one-year mortality of NYHA IV patients was 40% and the survival rate at 5 years was 25% for men and 38% for women [2]. The severity of HF is also marked by repeated unplanned hospitalizations due to Acute Decompensated HF (ADHF). Hospitalizations are frequent in HF patients and are associated with mortality and considerable economic burden. Indeed, there are around 150,000 admissions per year in France with more than 50% of HF patients requiring multiple hospitalization. These acute episodes have a poor prognosis with an evaluated risk of death of 40% within the following year. To improve the prognosis, an optimal management of hospitalized patients is necessary, but the ambulatory home follow-up has also a major impact. The current medical challenge for reducing these frequent hospitalizations remains in the early detection of ADHF at a pre-hospitalization stage. This would save valuable time and help to provide appropriate medical treatment to the patient.

One product developed to meet this clinical need based on a single direct biomarker is the Cardiomems (Abbott - St Jude Medical). This compact device is implanted in a branch of the pulmonary artery allowing a direct measurement of the pulmonary arterial pressure. The initial set of data collected from the use of this device showed that there was a reduction in re-hospitalization rate by 30% [3]. However, this device faces some major challenges and its stable acceptance in the market is limited by the long-term anti-aggregation therapies that are needed post implantation and its high cost (around 15,000€) [4].

Other diagnostic tools have been explored based on indirect biomarkers, such as body weight changes, bioimpedance or activity measurements [5, 6]. Although these parameters are useful, it has been established that a single sensor is clinically insufficient due to a lack of specificity and sensitivity. Cardiac Implantable Electronic Devices (CIED) allow the simultaneous use of multiple sensors for the recording of several indirect biomarkers. HeartLogic®, a diagnostic service compatible with Boston Scientific CIED triggers an alert of HF-worsening. The detection sensibility is 70% with a median alert window of 34 days [7]. However, despite the clinical benefit in the early detection of HF-worsening, only a small proportion of HF patients are eligible for an implantation with CIED.

Besides, other approaches have been considered such as band aids, patches or jackets, which are alternative remote monitoring solutions, but the effectiveness of these multiparametric approaches is conditioned by the patient's acceptance to permanently wear these medical devices which generally induce serious side effects on the skin, and consequently, a poor patient compliance.

Following this promising approach for very early intervention before ADHF, the strategy described hereafter is to provide a remote monitoring system for the integrated management of HF through a minimally-invasive device implanted into the stomach, more precisely in the fundus which is located a few millimeters from the heart, capable of recording multimodal parameters. The multimodal nature of the system could allow a reliable and early detection of ADHF (both sensitive and specific). From several sensors, various parameters will be recorded and used to build a predictive composite index of ADHF. Examples include heart rate (HRV) and respiratory rate variability, heart (S1, S3) and pulmonary sounds as well as patient activity and position. The composite index will be based on the variation of these different parameters and will allow to evaluate the decompensation state of each patients. In case of prediction of ADHF, an alarm will be sent to start an

appropriate medical care and avoid hospitalization or at least complications. The originality of our approach lies in the choice of a gastric implantation site [8]. This implantation site has never been used before and opens a promising exploratory field associated with new potentially relevant parameters in the follow-up of HF. The added value of the gastric implantation site also lies in the compatibility with an endoscopic route of introduction, thus minimizing complications and time of hospitalization, and in the patient compliance.

This article presents the first proof of concept obtained experimentally in pigs. A first prototype was developed to allow the acquisition of electrophysiological and mechanical signals in an in-vivo ambulatory context.

Methods

Prototype description

The prototype is composed of two parts connected with a 80 cm flexible cable, a gastric capsule (Figure 1) and an external module.

The gastric capsule is 30 mm long, 9 mm wide and 7 mm high. The body of the capsule is made with polyetheretherketone (PEEK), a biocompatible material resistant to hydrochloric acid. Each extremity corresponds to an electrode in titanium with a 35mm² surface and an inter-electrode distance of 20 mm. These electrodes are both connected to an electrocardiogram (ECG) chip (ADS1292, Texas Instrument, USA) embedded into the capsule. An anchorage ring is located at the end of one of the electrodes to allow the capsule to be secured by a stitch. A 3D accelerometer (ADXL355, Analog Device, USA) is also embedded in order to measure cardio-respiratory activity. The external module contains the batteries and a Bluetooth chip (BGM113, Silicon Labs, USA) for data transmission from the gastric capsule to the computer. An ECG chip and a 3D accelerometer are also embedded in the external module for reference signal acquisition. The data acquired by sensors into the gastric capsule and the external module are transmitted via Bluetooth Low Energy to a gateway connected to a laptop and are saved on the hard drive. All acquisitions are led by the software running on the laptop.

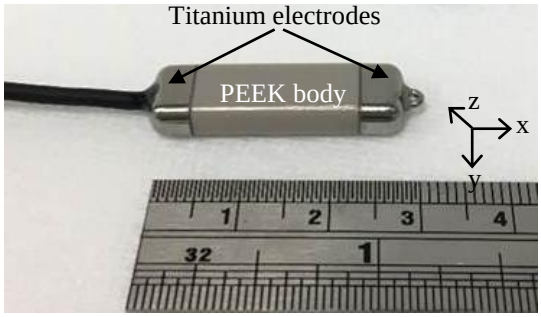


Figure 1: Gastric capsule design – the three orthogonal axis system represents the orientation of the accelerometer inside the gastric capsule.

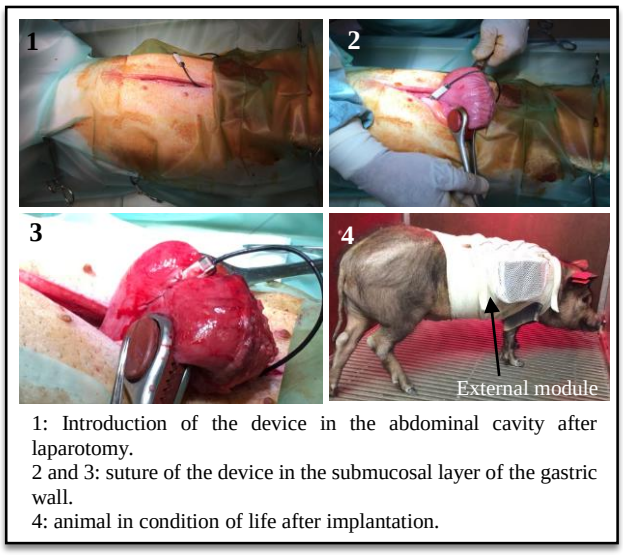
Animal experiments

Implantation procedure

In order to demonstrate the feasibility of our solution, 2 Yucatan male pigs weighing 21 kg were included in a first preliminary study. The protocol was approved by an Ethics Committee (n°C2EA-02) recognized by the French Ministry of Research. Each animal received humane care in accordance with European Directive 2010/63/EU on the protection of animals used for scientific purposes. Animals were fasted 24 hours before procedure. Prior to surgery, animals were pre-medicated by Ketamine, Azaperone, glycopyrronium bromide and Butorphanol administration. Propofol injection was used for anesthesia induction and Isoflurane (1-3%) was continuously administered through tracheal intubation for anesthesia maintenance. Heart rate, blood pressure, oxygen saturation and body temperature were monitored throughout the procedure.

The device implantation started with a median laparotomy, then a punctiforme incision was performed in the left flank. The device was introduced in the abdominal cavity through this incision. An incision of the serous and muscular layers at the level of the large gastric curvature of the stomach allowed to stitch the device in a submucosal location. The capsule was placed parallel to the anatomical axis of the heart to ensure optimal ECG measurement. Once the gastric capsule was implanted, it was connected to the external module and a first acquisition was performed to check the functionality and the positioning of the device. Then, the suture of the abdominal lining was completed (Figure 2).

After the end of surgery, pigs were slowly allowed to recover from anesthesia. Two adhesive external electrodes were placed, one on the anterior thorax and the other one on the posterior thorax, to allow acquisition of reference ECG. The external module was put in a pig jacket and pigs were returned to their cage (Figure 2). They received care every day during all the post-surgical follow-up of 1 month.



1: Introduction of the device in the abdominal cavity after laparotomy.
2 and 3: suture of the device in the submucosal layer of the gastric wall.
4: animal in condition of life after implantation.

Figure 2: Device implantation procedure

Signal acquisition

Acquisition was performed in automatic mode. ECG and accelerometric signals for both the gastric capsule and the external module were recorded during 30 seconds every 30 minutes during the day (8 am to 22 pm), and every 15 minutes during the night (22 pm to 8 am). This choice was based on the assumption that the recorded signals would be of better quality during the phases of inactivity and therefore by night.

The electrophysiological and mechanical parameters observed from the ECG and accelerometric signals are listed in Table 1.

Table 1: Parameters observed from raw signals recorded by the prototype

Raw signal	Parameters
ECG	Heart rate
	Heart Rate Variability
	P-wave
Accelerometer	S1
	S2
	Activity

Anatomopathological analysis

At the end of the post-surgical follow-up, the animals underwent an endoscopic examination of their stomach wall. Then, the system was removed using the same procedure as described above. Samples of the tissue surrounding the capsule were collected and directly immersed in a fixing solution for anatomopathological analysis.

Results

Anatomopathological analysis

The capsules were removed 1 month after implantation. The histological analysis of the samples of gastric tissue in direct contact with the capsule confirmed the location of the gastric device in the submucosal layer of the stomach and evidenced a moderate inflammatory and polymorphic macrophage reaction (Figure 3). Some fibrous changes were organized around this space but in limited abundance. No acute peritonitis was observed.

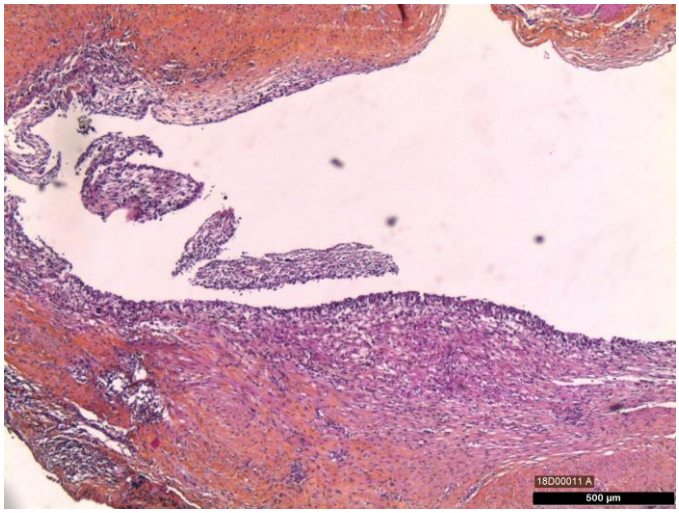


Figure 3: Cross section of the gastric wall at the implantation site - Hemalun Eosine staining

Parameter analysis

Following the implantation of the prototypes in the two pigs, ECG and accelerometric signals were recorded during 42 hours in the first pig and 14 days in the second one. After that time, it was impossible to communicate with the gastric capsules and no data could be recorded. Signals were processed using Matlab® to observe parameters of interest presented in Table 1.

ECG

Gastric ECG and external ECG were processed using the same algorithms. Figure 4a shows the raw gastric ECG. After filtering the ECG signal with a high-pass filter at 1 Hz to remove baseline variations (Figure 4b), R-peaks are clearly identifiable with an amplitude between 100 and 140 µV. The amplitude variation of R-peaks at low frequency (around 0.2 Hz) is due to respiration.

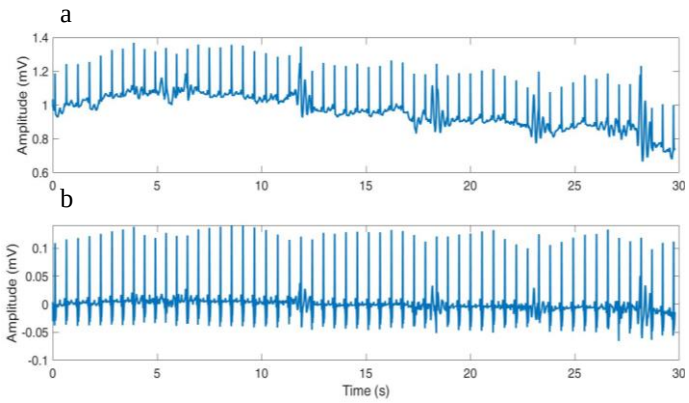


Figure 4: Gastric ECG - a) raw signal ; b) high-pass filtered signal

Once the gastric and external ECG signals were filtered, a R-peak detection was performed on a 30 second signal period using the Pan-Tompkins method [9]. The RR intervals in both gastric ECG and external ECG were calculated. There were no differences in the number of R-peaks detected. Figure 5 shows a dispersion of the dots of +/- 1 ms around the mean value due to the sampling frequency effect (1 kHz), as well as a measurement bias of 2 ms (Figure 5). This latter bias is related to internal clock discrepancies of the two ADS1292 devices used to perform the measurements.

After segmenting ECG signals into independent cardiac cycles, an ensemble average was performed from 15 consecutive cycles (Figure 6). A difference in amplitudes between the external (blue line) and gastric (red line) signals was observed (Figure 6a). The R-peak amplitude was 1.8 mV for the external ECG and 0.15 mV for the gastric ECG. This can be explained by the difference in electrode location and the inter-electrode distance. After amplitude normalization on the R-peak maximum value of the external and gastric ECG signals (Figure 6b), the shape of the two signals appears very similar. It can be highlighted that P-wave is clearly identifiable and more visible on the gastric signal than on the external one.

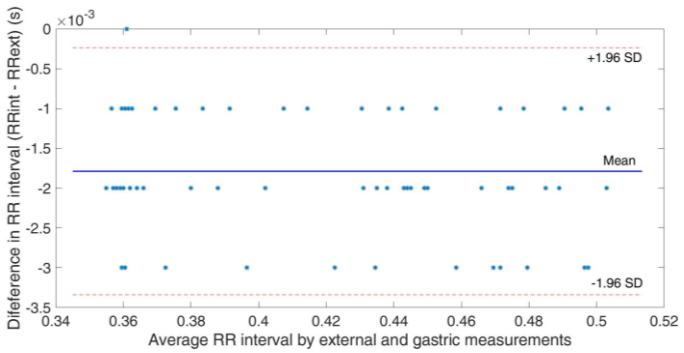


Figure 5: Bland & Altman plot – comparison of the RR intervals in both gastric and external ECG

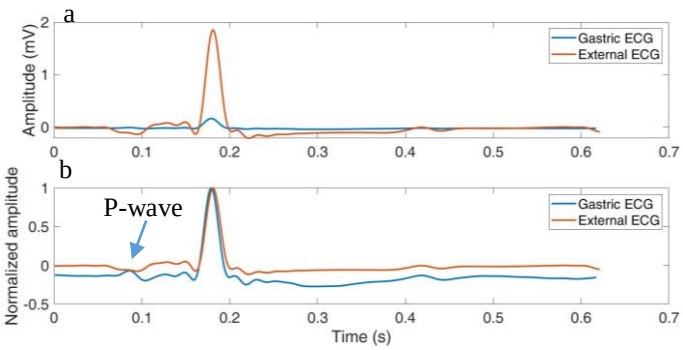


Figure 6: Ensemble average of both gastric and external ECG computed from 15 consecutive cycles – a) External ECG (red line) and gastric ECG (blue line) superimposed ; b) Normalized amplitude.

Accelerometric signals

Both accelerometric signals were processed to visualize the activity level on signal low-pass filter under 7 Hz. The gastric accelerometric signal was processed to extract: 1) the seismocardiographic signal on the extended band of 6 – 90 Hz; 2) the image of S1 and S2 heart sounds (accelerometric S1 and S2) on an audible band of 20 – 90 Hz. S1 and S2 are both caused by cardiac valve closure (mitral and tricuspid valves, aortic and pulmonary valves respectively) whose resulting vibrations are measured by the accelerometer [10].

In a first step, we calculated on 30 seconds epoch duration the three axis acceleration magnitude standard deviation to estimate a level of activity. Note that external measurements from the accelerometric sensor embedded in the external module only reflects the physical activity of the animal (no seismocardiographic component). From the external data, it is possible to separate activity and inactivity phases with an arbitrary threshold at 10 mg (Figure 7). During inactivity phases, the magnitude of signal perceived by the gastric accelerometer are over the level of the physical activity assessed by the external accelerometer. This level reflects additional mechanical activities (cardiac activity, respiration…). A first analysis revealed that the activity level impacts the quality of the signal.

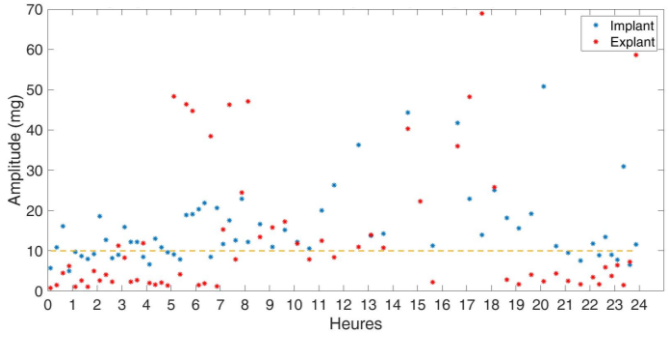


Figure 7: An exemple of an activity representation on a 24-hour period from gastric and external accelerometric measurements.

In the following, we only considered signals acquired during inactivity phases.

A first visual analyze of the 3 axes seemed to indicate that the Z axis carries the bulk of the signal. We therefore worked on the Z axis in the following. The signals on the audible band have visible and reproducible patterns at each cardiac cycle, corresponding to accelerometric S1 and S2 (Figure 8).

Figure 9 shows accelerometric signal ensemble average. The peak to peak signal amplitude is about +/- 10 mg in the extended band (Figure 9a) and +/-4 mg in the audible band (Figure 9b).

The residual noise observed on the audible band is comparable to the noise induced by the ADXL355 accelerometrer (0.2mg in the considered frequency band).

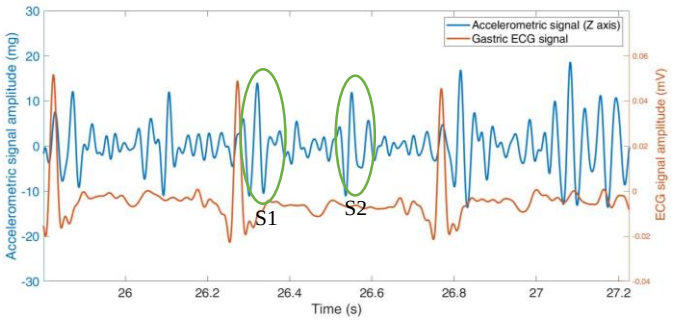


Figure 8: Gastric ECG and band-pass filtered (20-90 Hz) accelerometric signal - 3 consecutive cardiac cycles.

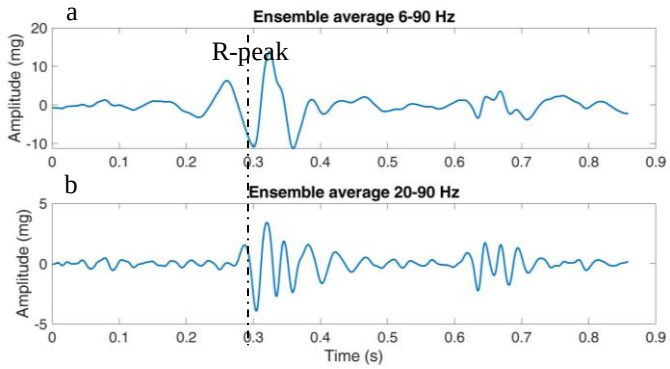


Figure 9: Z-axis ensemble average computed from 50 consecutive cardiac cycles a) in extended band (6 – 90 Hz); b) in audible band (20 – 90 Hz).

To confirm that the Z-axis is the best axis to specifically analyze cardiac activity, we compared the results to those obtained on the axis that represents the largest signal energy calculated by Principal Component Analysis (PCA) on the three axes. In figure 10, we compare the ensemble average obtained from the axis identified by the PCA method and the ensemble average from the Z axis at different times of acquisition. In Figure 10a, the Z-axis presents a greater signal to noise ration than PCA on S1. Figure 10b shows no significant differences between the Z-axis and PCA.

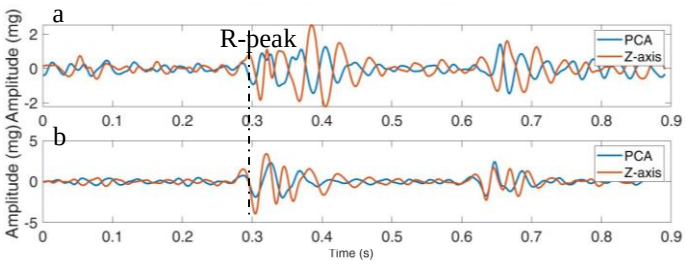


Figure 10: Comparison between the ensemble average performed on the Z-axis and on the axis identified by PCA – a) Extended band (6 – 90 Hz); b) Audible band (20 – 90 Hz).

Discussion

Electrophysiological and mechanical signals were recorded during two weeks from a capsule implanted in the stomach of two pigs. Gastric ECG signals were very similar to the reference external ECG. The P-wave representing atrial depolarization was more visible on the gastric ECG. This might be due to a closer location of the gastric capsule to the atrial site. In human, the fundus of the stomach is located at the level of the atrioventricular septum. These results should therefore be

transposable to humans and validate the positioning of the capsule.

Signal quality, both in ECG and accelerometric signals, is correlated to the physical activity level of the animal which is a good indicator to target exploitable signals. In this preliminary analysis, we only considered signals acquired during phases of inactivity. Indeed, the signals with an activity level above 10 mg present a greater noise level which is not completely removed by the ensemble average, and require a specific signal processing to analyze heart sounds. Physical activity might lead to resonance phenomena related to the deformation of the internal anatomical structures that are in the same bandwidth as the heart sounds making their analysis difficult. However, noise conditions are not stationary even during inactivity phases and must be analyzed in regard with the physiological conditions in order to understand the difference in activity level and its impact on noise level. The signal amplitude measured by the gastric accelerometric sensor is over the physical activity level measured by the external accelerometer sensor which is decoupled from seismocardiographic activity. This might be related to the mechanical cardiac activity or other anatomical structure movements. According to the literature, the amplitude of cardiac activity signal on the extended band acquired from the thorax is +/-8mg, with a noise of +/-3mg [11], which is consistent with our results during phases of inactivity. Comparison between Z axis and PCA did not exhibit reorientation problem. However, statistical analysis must be conducted to confirm this results in different conditions of noise. Although, accelerometric S1 and S2 can be visualized from signal during inactivity phases, the challenge is now to detect the accelerometric S3 in pathological conditions. Let us remind that S3 is specific of cardiac decompensation. Further studies are required to assess the statistical value of our results and to optimize the signal processing methods in order to extract heart sounds and other parameters of interest from the accelerometric signal. In this context we now develop multimodal methods which could be relevant in such noisy conditions.

Conclusions

This proof of concept in an in-vivo ambulatory context confirms our capacity to identify some phases on which electrophysiological and mechanical parameters of interest can be observed with a device implanted in the stomach wall. However, noise conditions must be explored further to optimize and adapt the signal processing method that will allow to increase the number of exploitable acquisition phases (activity…).

To further demonstrate the feasibility of monitoring variations of cardiac and respiratory parameters in patients with HF, experiments are in progress in a pig model of HF (leading to ADHF) using this technology. For this purpose a fully implantable version of the device is currently developed for long term implantation.

Acknowledgements

The authors would like to thank the team of the Human Nutrition Unity (INRA - Clermont-Ferrand) for their involvement in animal experiments.

References

[1] T. Saudubray, C. Saudubray, C. Viboud, G. Jondeau, A. Valleron, A. Flahault, and T. Hanslik, Prevalence and management of heart failure in France: national study among general practitioners of the Sentinelles network, *La Revue de medecine interne* **26** (2005), 845-850.

[2] K.K. Ho, J.L. Pinsky, W.B. Kannel, and D. Levy, The epidemiology of heart failure: the Framingham Study, *Journal of the American College of Cardiology* **22** (1993), A6-A13.

[3] W.T. Abraham, P.B. Adamson, R.C. Bourge, M.F. Aaron, M.R. Costanzo, L.W. Stevenson, W. Strickland, S. Neelagaru, N. Raval, and S. Krueger, Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial, *The Lancet* **377** (2011), 658-666.

[4] A.T. Sandhu, J.D. Goldhaber-Fiebert, D.K. Owens, M.P. Turakhia, D.W. Kaiser, and P.A. Heidenreich, Cost-effectiveness of implantable pulmonary artery pressure monitoring in chronic heart failure, *JACC: Heart Failure* **4** (2016), 368-375.

[5] D.J. Whellan, K.T. Ousdigian, S.M. Al-Khatib, W. Pu, S. Sarkar, C.B. Porter, B.B. Pavri, C.M. O'Connor, and P.S. Investigators, Combined heart failure device diagnostics identify patients at higher risk of subsequent heart failure hospitalizations: results from PARTNERS HF (Program to Access and Review Trending Information and Evaluate Correlation to Symptoms in Patients With Heart Failure) study, *Journal of the American College of Cardiology* **55** (2010), 1803-1810.

[6] D. Slotwiner, N. Varma, J.G. Akar, G. Annas, M. Beardsall, R.I. Fogel, N.O. Galizio, T.V. Glotzer, R.A. Leahy, and C.J. Love, HRS Expert Consensus Statement on remote interrogation and monitoring for cardiovascular implantable electronic devices, *Heart Rhythm* **12** (2015), e69-e100.

[7] J.P. Boehmer, R. Hariharan, F.G. Devecchi, A.L. Smith, G. Molon, A. Capucci, Q. An, V. Averina, C.M. Stolen, and P.H. Thakur, A multisensor algorithm predicts heart failure events in patients with implanted devices: results from the MultiSENSE study, *JACC: Heart Failure* **5** (2017), 216-225.

[8] P. Cinquin, P. Defaye, F. Boucher, P.Y. Guméry, Intragastric measuring device, FR1662059, 7/12/16.

[9] Pan and W.J. Tompkins, A real-time QRS detection algorithm, *IEEE Trans. Biomed. Eng* **32** (1985), 230-236.

[10] P. Castiglioni, P. Meriggi, F. Rizzo, E. Vaini, A. Faini, G. Parati, G. Merati, and M. Di Rienzo, Cardiac sounds from a wearable device for sternal seismocardiography, *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, IEEE, 2011, pp. 4283-4286.

[11] K.Z. Siejko, P.H. Thakur, K. Maile, A. Patangay, and M.T. OLIVARI, Feasibility of heart sounds measurements from an accelerometer within an ICD pulse generator, *Pacing and Clinical Electrophysiology* **36** (2013), 334-346.

Address for correspondence

Cindy Dopierala – cindy.dopierala@univ-grenoble-alpes.com

Un modèle sémantique d'identification du médicament en France

J. Grosjean[1,2], C. Letord[1,2], J. Charlet[2,3], X. Aimé[2], L. Danès[2], J. Rio[1], I. Zana[2], SJ. Darmoni[1,2], C. Duclos[2,3]

[1] Department of Biomedical Informatics, Rouen University Hospital, 76031 Rouen Cedex, France;
{Julien.Grosjean,Catherine.Letord,Julien.Rio,Stefan.Darmoni}@chu-rouen.fr
[2] Sorbonne Université, INSERM, Université Paris 13, LIMICS, Paris, France;
xavier.aime@cogsonomy.fr,loane.danes@agroparistech.fr,ilan.zana26@gmail.com
[3] Assistance Publique-Hôpitaux de Paris, DRCI, Paris, France
{Jean.Charlet,Catherine.Duclos}@aphp.fr}

Résumé : il n'existe pas de standard universellement accepté pour nommer les médicaments. L'identification du médicament a fait l'objet de nombreux travaux de normalisation. Notre objectif est de définir un modèle formel du médicament en français pour lier les différentes entités manipulables autour du médicament. Ce modèle formel vise un double sous-objectif : (a) créer et instancier une ontologie formelle du médicament ; (b) créer une terminologie du médicament, intégrable dans un serveur de terminologies. À terme, ces ressources seront des outils puissants pour, notamment, supporter la recherche d'information dans des bases de médicaments ou des entrepôts de données. Ils seront mis librement à disposition de la communauté.

Mots-clés : ontologie ; terminologie ; serveur de terminologie ; modèle formel ; médicament.

1 Introduction

Bien que le médicament soit fini, identifiable, il n'existe pas de standard universellement accepté pour le représenter (Cimino 1999). Selon le point de vue auquel on se place, on peut le définir à un niveau moléculaire (comme une substance active), à un niveau clinique (comme un produit capable de traiter une pathologie) ou encore à un niveau physique (comme une présentation destinée à satisfaire la prescription et délivrable au patient). L'identification du médicament peut donc se concevoir à divers niveaux ayant des degrés d'abstraction plus ou moins grands (Sperzel 1998) (i) une présentation, un produit manufacturé ou un ingrédient correspondent à des objets physiques, (ii) un produit clinique ou une fraction thérapeutique sont de pures abstractions.

L'identification du médicament a fait l'objet de nombreux travaux de normalisation dont les plus récents définissent l'identification du produit médicinal et du produit pharmaceutique (ISO 11615, ISO 11616, ISO 20443, ISO 11238, ISO 11239, ISO 11240) afin de rendre le partage international de l'information sur le médicament possible. Le référentiel résultant sera disponible en 2020 au niveau de l'agence européenne du médicament (EMA). D'autres modèles ont été adoptés pour représenter le médicament dans les bases de données sur le médicament (Broverman 1998, DMD[1]), ou pour servir de pivot entre des bases des données sur le médicament (RxNorm[2]). Dans ces modèles, se retrouve cette dualité entre virtuel et réalité et les relations de composition entre ingrédients actifs, dosages et formes.

[1] https://apps.nhsbsa.nhs.uk/DMDBrowser/DMDBrowser.do
[2] https://www.nlm.nih.gov/research/umls/rxnorm

Ces référentiels « normalisés », lorsqu'ils sont disponibles, n'incluent pas de médicaments français. L'agence nationale de sécurité du médicament et des produits de santé (ANSM) met à disposition, via la Base de Données Publique du Médicament (BDPM[3]), des fichiers décrivant les médicaments français et leur composition mais ne respectent pas toujours les normes d'identification du médicament qui en permettraient un usage simple. Une équipe bordelaise a proposé une transformation de ces fichiers dans le format RxNorm afin de disposer de nombreux variants dénommant les médicaments pour rechercher ces derniers dans des comptes rendus textuels de passage aux urgences (Cossin 2018).

Dans ce travail, nous proposons et présentons un modèle formel du médicament en français pour lier les différentes entités manipulables autour du médicament. Ce modèle sera disponible gratuitement pour la communauté scientifique. Ce modèle formel vise un double sous-objectif : (a) créer et instancier une ontologie formelle du médicament, en s'appuyant sur des données librement accessibles, en particulier celles fournies par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé (ANSM) via la BDPM et celles issues des bases de données bibliographiques, comme PubMed ou LiSSa ; (b) créer une terminologie du médicament, intégrable dans un serveur de terminologie. Ce modèle servira en première intention au projet PsyHAMM, dont le but est de détecter des prescriptions hors AMM (Autorisation de Mise sur le Marché) en psychiatrie[4]. Idéalement, ce modèle pourra également être utilisé pour rechercher une information de qualité sur les médicaments dans les différents entrepôts de données développés en France. Une représentation formelle et donc standardisée permet en outre de mieux contrôler la qualité des informations autour du médicament (typiquement via la détection des inconsistances d'une ontologie).

Dans la prochaine section nous décrivons toutes les actions de normalisations et les difficultés rencontrées. Dans la section 3, nous décrivons la méthodologie de construction des modèles et nous donnons un certain nombre de résultats dans la section 4. Nous terminons par quelques perspectives en section 5.

2 Définitions, référentiels et informations sur le médicament

L'analyse des modèles existants permet de repérer les concepts retrouvés pour identifier le médicament, à savoir la substance active, l'excipient, la dose, la forme, la voie d'administration, le nom commercial, le conditionnement, et la combinaison de ces concepts permettant de définir des produit cliniques, pharmaceutiques, médicinaux, et présentés. Ces concepts peuvent être décrits dans des terminologies comme l'ATC[5], le MeSH[6], la SNOMED[7] ou encore l'UMLS. Des référentiels normatifs existent aussi pour représenter les formes, voies d'administration (Standard Terms de l'EMA[8]), les unités (UCUM). Par ailleurs une norme française d'interopérabilité (NF 97-555) définit en France des identifiants de médicaments (CIS, UCD, CIP), les liens entre eux et l'attendu quant à la construction de la dénomination d'une spécialité pharmaceutique.

La Banque Publique du Médicament met à disposition les fichiers (a) de spécialités pharmaceutiques associant un identifiant CIS à une dénomination (devant normalement être construite selon le modèle nom de marque, dosage, forme), (b) de composition permettant de connaître la composition de la spécialité pharmaceutique en quantité de substance active et en fraction thérapeutique, (c) des présentations associant un/des identifiant(s) CIP à une/plusieurs présentations elles même reliées à une spécialité pharmaceutique. Ces informations sont associables à un résumé des caractéristiques du produit (RCP). Le RCP est une annexe de la

[3] http://base-donnees-publique.medicaments.gouv.fr/

[4] https://anr.fr/Projet-ANR-18-CE19-0017

[5] https://www.whocc.no/atc/

[6] http://www.nlm.nih.gov/mesh

[7] https://www.health.belgium.be/fr/terminologie-et-systemes-de-codes-snomed-ct

[8] https://www.edqm.eu/en/standard-terms-database

décision d'autorisation synthétisant les informations notamment sur les indications thérapeutiques, contre-indications, modalités d'utilisation et les effets indésirables d'un médicament. Par exemple, « PROZAC 20 mg, gélule » est le libellé d'une spécialité pharmaceutique qui a pour code CIS 61885224. Elle contient de la « fluoxétine » (substance) sous forme de « chlorhydrate de fluoxétine » (un sel possible de cette substance). Sa vente est autorisée sous la présentation « plaquette(s) thermoformée(s) PVC-Aluminium de 14 gélules ». Cette présentation est référencée par un code identifiant de présentation (CIP) à 13 chiffres figurant sur la boîte du médicament.

Enfin il existe un identifiant (code UCD) correspondant au plus petit élément commun à plusieurs présentations d'une même spécialité pharmaceutique. Le code UCD représente, pour chaque forme galénique, la plus petite unité de dispensation (comprimé, flacon, ...). Le code UCD et le nombre d'UCD par présentation sont administrés par le Club Inter Pharmaceutique[9].

La description du médicament par la BDMP est limitée par rapport aux attendus des différents modèles d'identification du médicament. Par exemple, le terme « PROZAC » n'est pas présent dans cette base, alors que « PROZAC 20 mg, gélule » l'est. Nous avons créé des racines pharmaceutiques correspondant aux différents noms commerciaux (ici « PROZAC ») présents de la BDPM. Ces ajouts permettent d'améliorer le rappel quand on applique un annotateur sémantique pour la détection des médicaments dans un document de santé. Cela permet également de regrouper les CIS ayant la même racine.

Plus récemment, quatre bases de données médicamenteuses labellisées par la Haute Autorité de Santé (HAS) se sont associées pour créer Medicabase[10], une base de données de médicaments virtuels. Les médicaments virtuels permettent de regrouper des spécialités qui comportent :

- le ou les même(s) principe(s) actif(s) ou des sels du ou des principe(s) actif(s) cliniquement équivalent(s) du point de vue des risques iatrogènes ;
- les mêmes dosages en base active des principes actifs ;
- une forme galénique considérée comme cliniquement équivalente du point de vue des risques iatrogènes.

Par exemple, l'« Abacavir 300 mg comprimé » regroupe l'« ABACAVIR MYLAN 300 mg, comprimé pelliculé sécable » et l'« ABACAVIR SANDOZ 300 mg, comprimé pelliculé sécable », qui sont des spécialités équivalentes.

Par ailleurs, au sein des fichiers de la BDPM, les voies d'administration semblent normalisées, au contraire des formes pharmaceutiques qui ne semblent pas respecter un modèle précis, encore moins les conditionnements (Cf. *infra*).

Le D2IM travaille depuis 10 ans (Pereira, 2008) sur le médicament afin d'intégrer le plus d'informations structurées concernant ce dernier au sein de son serveur terminologique HeTOP (Grosjean, 2012). Mais, il manquait à ce travail un modèle formel permettant par exemple, de retrouver tous les codes CIP ou UCD pour une substance active donnée ou pour un médicament virtuel donné. Pour finir, le LIMICS a initialisé en 2016 une ontologie sur le médicament mais le projet s'est heurté à de nombreux problèmes de cohérences et de qualité des bases utilisées pour construire cette ontologie (Steinberg 2016).

3 Processus de construction de l'ontologie et de la terminologie des médicaments

À partir du travail des équipes impliquées et au regard des problèmes de qualité des fichiers d'origine, nous avons utilisé le travail ontologique comme un test de la cohérence des données issues de la BDPM et nous avons mis au point un processus de construction d'une base terminologique et d'une ontologie qui se déroule en plusieurs étapes avec un outil d'ETL (Talend) appliqué aux fichiers disponibles :

1. création d'une ontologie de départ avec un certain nombre de classes normalisées :

[9] http://www.ucdcip.org/

[10] http://www.medicabase.fr/

   a. intégration de l'ATC,
   b. hiérarchie des voies d'administration développées par un pharmacien,
   c. création des classes correspondant aux conditions d'utilisation des médicaments (liste à faire évoluer en fonction des évolutions (rares) de ces conditions),
   d. création des classes propres au modèle du médicament de l'ANSM (nom commercial, substance active, générique, …)

2. récupération des fichiers publics issus de l'ANSM, de Medicabase, et de l'ATC. Certains codes ATC sont corrigés en avance de phase par le D2IM pour les médicaments les plus récents. La table 1 résume l'ensemble des fichiers utilisés ;
3. vérifications de la cohérence des fichiers avec des tests. La table 2 liste les principaux tests de cohérence réalisés ;
4. création de l'ontologie de travail instanciant le modèle de l'ontologie de départ avec les fichiers décrits plus haut, en incluant la création de hiérarchies, d'une part pour les conditionnements et d'autre part pour les formes ; les deux étant issus des fichiers de l'ANSM ;
5. fourniture de l'ontologie de travail pour être intégrée et maintenue (processus manuel d'assurance qualité et processus automatique de contrôles d'intégrités) au sein du serveur terminologique d'HeTOP ; les différentes métadonnées du modèle instancié sont transposées dans une base de données relationnelle et l'intégrité référentielle contrôle les éventuels codes erronés.
6. récupération de fichiers corrigés issus de HeTOP et génération de l'ontologie finale.

Tableau 1 – Liste des principaux fichiers utilisés pour l'extraction des données sur le médicament

CIS.txt	Donne la liste des spécialités et de leurs détails.
COMPO.txt	Donne la liste des molécules, substance active ou fraction thérapeutique, et les spécialités qu'elles composent.
CIS_CIP.txt	Donne la liste des présentations (codes CIP) correspondant à chaque spécialité (code CIS) et leurs détails.
cis_cpd_bdpm.txt	Donne pour chaque spécialité la condition d'utilisation.
cis_gener_bdpm.txt	Donne pour chaque générique, la ou les catégories auxquelles il appartient, sa dénomination et sa/ses spécialités correspondantes.
Autorisations actives - ATC_mdts présents sur RSP.xls	Pour tous les médicaments possédant une AMM donne la relation entre le code ATC et chaque spécialité.
fic01den.txt	Donne la liste des groupes génériques et des DIC correspondantes.
fic02grp.txt	Donne la liste des génériques, le groupe auquel chacun appartient, sa voie d'administration et sa dénomination.
fic03spe.txt	Donne pour chaque spécialité, la catégorie à laquelle elle appartient, son générique correspondant et sa dénomination.
liste_medicamentVirtuels	Donne la liste des médicaments virtuels.
ATC_2018.owl	Donne l'arborescence ATC sous forme de fichier OWL avec les labels normalisés.

4 Résultats

Le modèle d'identification du médicament en français est détaillé dans la Figure 1. Nous avons travaillé à sa création, avec dès le départ, un test sur plusieurs cas d'usage, en démarrant

par les plus simples (en évitant pour l'instant les associations médicamenteuses complexes et les médicaments biologiques). Ce modèle tente de minimiser les relations entre les différents concepts du médicament. Le reste des relations se calculent, comme par exemple la relation entre un médicament virtuel et un code ATC se déduit par les deux relations entre médicament virtuel et spécialité pharmaceutique d'une part, et spécialité pharmaceutique et code ATC d'autre part (existant auparavant dans la base, sur un ancien modèle).

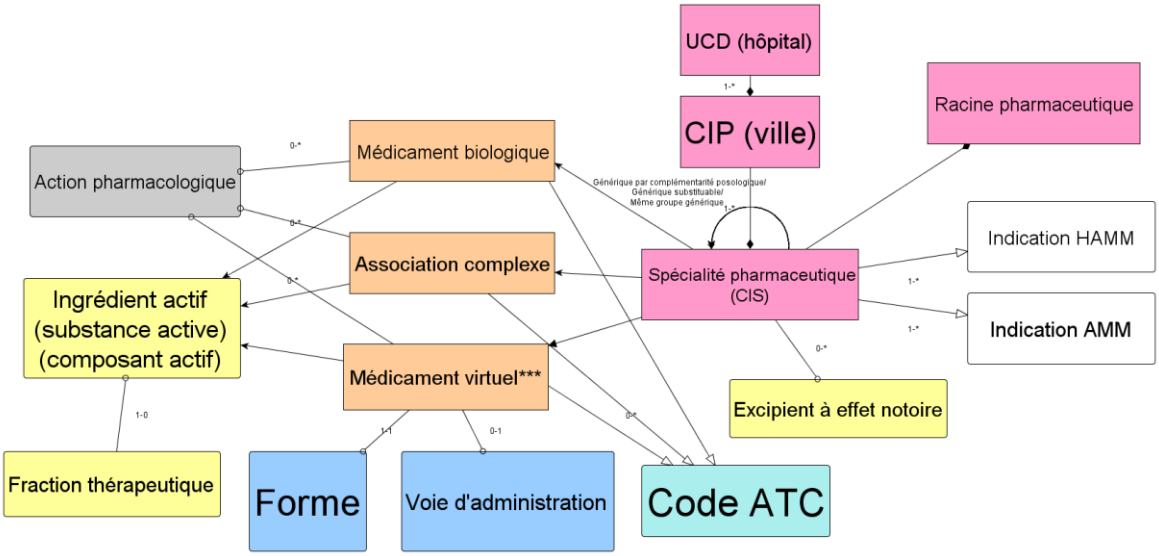


Figure 1 : modèle sémantique du médicament en France. La plupart de ces entités correspondent déjà à des ensembles de codes ou nomenclatures qu'il est nécessaire de mettre en correspondance. Les liens sémantiques sont définis explicitement ainsi que les cardinalités associées. Ces liens sont définis de la façon la "plus haute" possible afin de ne pas répéter l'information. Ce modèle est à la fois instancié sous forme de terminologie et d'ontologie pour des utilisations variées.

Ce modèle permet de retrouver un médicament dans un entrepôt de données, sous n'importe laquelle de ses formes : qu'il s'agisse de son nom commercial ou de ses ingrédients, son code ATC ou ses codes CIS, CIP ou UCD. Par exemple, pour la substance active (ou ingrédient) modafinil, le modèle nous fournit les informations suivantes : nom commercial = MODAFINIL; code ATC = N06BA07 ; quatre codes UCD (correspondant à quatre génériques) = 3400893594582, 3400893615676, 3400893926680, 3400894100188 ; le seul médicament virtuel actuellement est : Modafinil 100 mg comprimé. De plus, en utilisant HeTOP, nous pouvons remarquer que cette substance active existe dans les termino-ontologies suivantes : MeSH, NCIT (US National Cancer Institute), SNOMED CT, LOINC (Logical Observations Identifiers Names and Codes), BNPC (Base Nationale (Française) des Produits et des Compositions).

À partir de l'ontologie de travail, nous avons réalisé huit principaux tests de cohérence, dont le détail est fourni dans la table 2. Deux tests de cohérence ont détecté de nombreuses erreurs : 6 554 codes CIS n'ont pas de code CIP ; 509 codes CIS n'ont pas de codes ATC.

Tableau 2 - Principaux tests de cohérence et fichiers de contrôle générés, associés au nombre d'occurrences en janvier 2019.

Unfound_CIS_Compo.xls	Recense les codes CIS présents dans le fichier CIS.txt mais absents du fichier COMPO.txt.	2
Unfound_CIS_CisCip.xls	Recense les codes CIS présents dans le fichier CIS.txt mais absents du fichier CIS_CIP.txt.	5

Without_Libelle_Cip13.xls	Recense les codes CIS du fichier CIS_CIP.txt qui n'ont pas de code CIP13 et/ou de libellé.	6554
Unfound_CIS_ATC.xls	Recense les médicaments associés à un code CIS dans le fichier CIS.txt qui ne sont pas présents dans le fichier ATC et qui donc n'ont pas de code ATC associé.	509
Unfound_ATC_Onto.xls	Recense les codes ATC associés à des codes CIS qui ne sont pas présents dans l'arborescence de l'ontologie (même après mise à jour).	15
Unfound_Fic3_Gener.xls	Recense la liste des codes CIS qui sont présents dans les fichiers des spécialités fic03spe.txt mais qui ne sont pas présents dans le fichier *cis_gener_bdpm.txt*.	118
Unfound_Fic3_CIS.xls	Recense la liste des codes CIS qui sont présents dans les fichiers des spécialités *fic03spe.txt*.	40
Voies_manquantes.txt (resp. voies_manquantes_fic.txt)	liste des voies qui sont présentes dans le fichier *CIS.txt* (resp. *fic02grp.txt*) mais qui sont absentes de l'ontologie.	1

Les conditionnements ne sont pas normalisés dans les fichiers de la BDPM. Le travail de l'ETL sur les conditionnements permet de normaliser les chaînes de caractère puis créer une hiérarchie. Ces conditionnements sont enregistrés sous formes de libellés associés aux classes en relations d'hyperonymie. Par exemple, on a une classe dont le libellé est « boîte », elle subsume une classe dont le libellé est « boîte aluminium ». Cette dernière subsume une classe dont le libellé est « boîte aluminium comprimé ». Enfin, on trouve en dessous les classes correspondant à six conditionnements déclarés dans les fichiers, « 1 boîte aluminium de 6/10/12/20/25/1000 comprimés ». En février 2019, il y a 25 661 chaînes de caractères différentes décrivant des conditionnements qui correspondent à autant de classes organisées en une hiérarchie dont le premier niveau comporte 99 classes et qui a une profondeur moyenne de 9.

Les formes ne sont pas non plus normalisées. Un travail du même type que les conditionnements a amené l'explicitation de 597 formes organisées en une hiérarchie dont le premier niveau comporte 100 classes et qui a une profondeur moyenne de 3,3. Il existe une liste anglo-saxonne de formes sur laquelle nous allons travailler pour améliorer et normaliser cette hiérarchie.

Une instanciation d'HeTOP a été développée pour afficher et synthétiser l'ensemble des informations concernant le médicament en se fondant sur le modèle présenté dans ce travail. Ce serveur « HeTOP médicament » est mis à disposition des utilisateurs qui ne sont pas experts du médicament dans une version compacte et simplifiée (https://www.hetop.eu/hetop/medicaments).

5 Discussion

Notre travail a démontré qu'il était possible à partir des fichiers de l'ANSM de construire un référentiel ontologique et terminologique moyennant la correction de nombreuses erreurs ou oublis qui sont étonnantes puisque ces fichiers doivent se conformer à une norme et sont issus d'une autorité de référence.

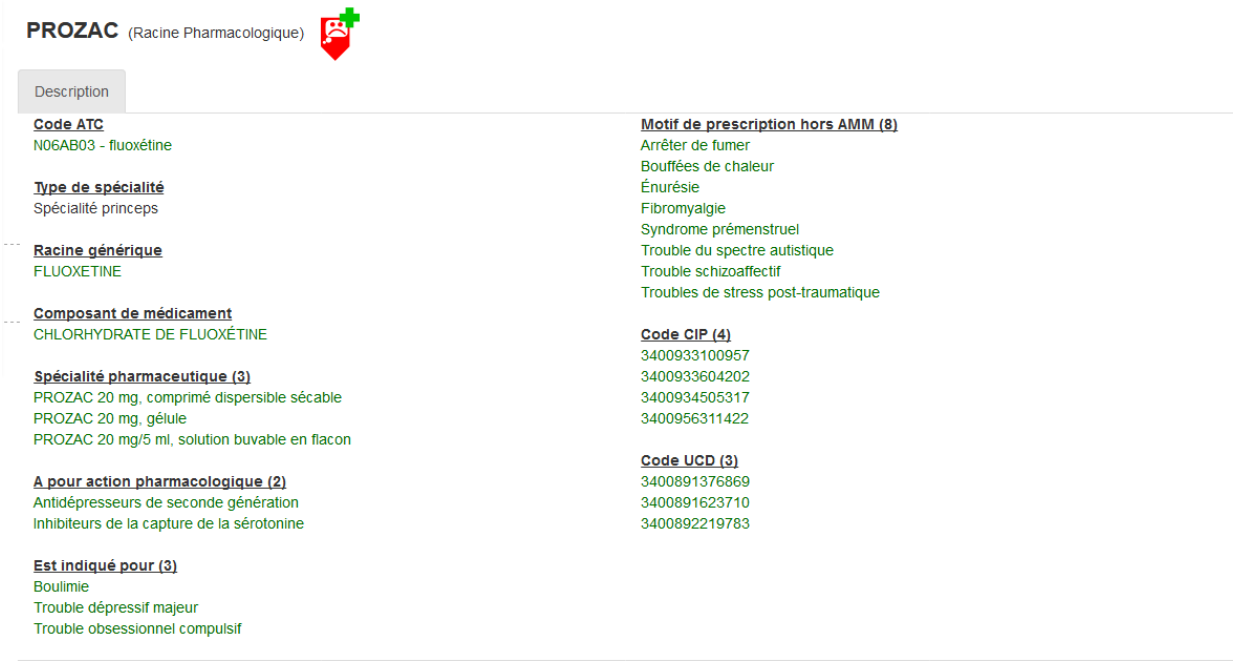


Figure 2 : exemple des informations sur un médicament fournies par le serveur terminologique « HeTOP médicaments » sur ECMT. Il s'agit d'une vue synthétique créée dynamiquement à travers tous les liens disponibles de façon profonde dans la terminologie produite dans cette étude.

Le développement d'un modèle ontologique et terminologique de façon coordonnée permet d'avoir deux versions des mêmes connaissances. Le couplage termino-ontologique permet de créer un cycle de qualité, en détectant certaines incohérences. La version terminologique, incluse dans HeTOP pourra être immédiatement évaluée dans le contexte de l'annotateur sémantique ECMT (Cabot 2016) et dans une exploitation de l'Entrepôt de Données de Santé du CHU de Rouen pour la recherche d'information sémantique (Ndangang 2018) La version ontologique sera utilisée dans un annotateur sémantique du LIMICS, d'abord dans le contexte du projet PARON (Cardoso 2018). Le format ontologique permettra de mettre à disposition des versions réduites de l'ontologie en fonction des contextes d'usage : il suffira de préparer les versions nécessaires avec des requêtes SPARQL.

Un projet proche est le projet ROMEDI (Cossin 2018), qui vise la détection de médicaments en texte libre. L'équipe de Bordeaux a produit un site Web[11], qui fournit de nombreuses informations intéressantes sur le médicament et qui permet maintenant de récupérer une ontologie représentant un certain nombre d'informations du site. Dès qu'elles seront bien stabilisées, les ressources mises à disposition par les deux projets pourraient être comparées. À l'heure actuelle, la présente approche ajoute un certain nombre d'informations d'intérêt comme les CIP, les UCD, les indications ou encore les CIS dont la commercialisation a été stoppée. Ces données sont cruciales pour une utilisation en vie réelle, dans les hôpitaux notamment.

Par la suite, un certain nombre d'étapes vont suivre :

- Notre modèle doit, en premier lieu, être validé par des médecins et des pharmaciens du consortium PSYHAMM n'ayant pas participé à sa mise au point. En particulier, les cas difficiles, comme les associations complexes et les médicaments biologiques, seront étudiés en priorité.
- À notre connaissance, les médicaments en Europe ne seront définis selon la nouvelle norme IDPM qu'en 2020. De nouveaux identifiants seront fournis par IDPM et seront aisément intégrables dans notre modèle formel.

[11] http://www.romedi.fr/

- Concernant les indications, il existe des informations en texte libre dans la BDPM et structurées dans les bases de données françaises labellisées par la HAS. L'intégration de ces informations au sein de la base terminologique puis de l'ontologie est au programme de l'équipe dans le cadre du projet PsyHAMM. On voit dans la figure 2, l'affichage d'informations importantes pour le projet, à savoir les motifs de prescriptions hors AMM. Ceux-ci seront retravaillés et complétés.

En conclusion, notre consortium a réalisé un modèle formel du médicament, avec un couplage termino-ontologique permettant un cycle de qualité. Les erreurs détectées dans les données sources seront remontées à leurs éditeurs respectifs et les bases de connaissances construites seront très bientôt diffusée librement à la communauté (en plus du site et service web de HeTOP). Les retombées de ce modèle sont nombreuses : en premier lieu, la possibilité de rechercher une information sur le médicament dans les entrepôts de données de santé en France, cette information pouvant être exprimée de multiples façons, grâce à une vision multi-terminologique.

Remerciements

Ce travail a été réalisé dans le cadre du projet ANR PSYHAMM (18-CE19-0017) financé par l'Agence Nationale de la Recherche et du projet FEDER PlaIR2.018 (Région Normandie).

Références

Broverman C, Kapusnik-Uner J, Shalaby J, Sperzel D. A concept-based medication vocabulary: an essential requirement for pharmacy decision support.Pharm Pract Manag Q. (1998) Apr;18(1):1-20.

Cabot C, Lelong R, Grosjean J, Soualmia LF & Darmoni SJ. Retrieving Clinical and Omic Data from Electronic Health Records.. Stud Health Technol Inform School. (2016):221, 115.

Cardoso S, Aimé X, Meininger V, Grabli D, Melo Mora LF, Bretonnel CK & Charlet J. A Modular Ontology for Modeling Service Provision in a Communication Network for Coordination of Care, Stud. Health Technol. Inform. (2018) 890–894. doi:10.3233/978-1-61499-852-5-890.

Cimino J, McNamara T, Meredith T, Broverman C, Eckert K, Moore M, et al. Evaluation of a proposed method for representing drug terminology. Proc AMIA annu Fall Symp. (1999) :47-51.

Cossin S, Loustau R, Jouhet V, Létinier L, Mougin F, Evrard G, Gil-Jardiné C, Diallo G & Thiessard F. ROMEDI, une terminologie médicale française pour la détection des médicaments en texte libre. (2018)

Bouhaddou O, Warnekar P, Parrish F, Do N, Mandel J, Kilbourne J & Lincoln MJ. Exchange of computable patient data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): terminology mediation strategy. J Am Med Inform Assoc. (2008) Mar-Apr;15(2):174-83.

Grosjean J, Merabti T, Griffon N, Dahamna B & Darmoni SJ. Teaching medicine with a terminology/ontology portal. Stud Health Technol Inform. (2012);180:949-53.

Ndangang M, Grosjean J, Lelong R, Dahamna B, Kergourlay I, Griffon N & Darmoni SJ. Terminology Coverage from Semantic Annotated Health Documents. Studies in Health Technology and Informatics (2018);255:20-24.

Pereira S, Plaisantin B, Korchi M, Rozanes N, Serrot E, Joubert M & Darmoni SJ. Automatic construction of dictionnaries, application to product characteristics indexing. Stud Health Technol Inform. (2009);150: 512-6.

Sperzel WD, Broverman CA, Kapusnik-Uner JE & Schlesinger JM. The need for a concept-based medication vocabulary as an enabling infrastructure in health informatics. Proc AMIA Symp. (1998):865-9.

Steinberg K. Qualité des données de santé disponibles en France et de leurs modèles - Comment la garantir pour répondre aux enjeux de la gestion des connaissances médicales ? Mémoire CNAM, Titre professionnel niveau 1 Chef de projet en ingénierie documentaire et gestion des connaissances. (2016). Disponible à http://portaildoc-intd.cnam.fr/Record.htm?idlist=1&record=19298817124910160999

EzMedRec: Une aide à la conciliation médicamenteuse sémantiquement enrichie

Brigitte Séroussi[1,2], Mourad B. Ghomari[1], Isabelle Debrix[2], Gilles Guézennec[1] et Jacques Bouaud[3,1]

[1]Sorbonne Université, Université Paris 13, Sorbonne Paris Cité, Inserm, LIMICS, Paris, France,
[2]Assistance Publique-Hôpitaux de Paris, Hôpital Tenon, Paris, France,
[3]Assistance Publique-Hôpitaux de Paris, DRCI, Paris, France

Résumé : La conciliation médicamenteuse vise à prévenir les erreurs médicamenteuses aux points de transition des parcours de soins. Cette tâche est complexe, chronophage et exigeante sur le plan cognitif. Il s'agit en effet de faire la synthèse de tous les médicaments administrés à un patient donné par les différents professionnels engagés dans sa prise en charge et de vérifier qu'il n'y a pas d'interaction médicamenteuse pouvant donner lieu à des événements indésirables graves au moment d'ajouter de nouveaux médicaments au traitement courant. EzMedRec est un système d'aide à la conciliation médicamenteuse rétroactive à l'admission à l'hôpital. L'objectif est de vérifier la cohérence entre la liste des médicaments pris par le patient avant l'hospitalisation, ou bilan médicamenteux optimisé (BMO), aux prescriptions établies à l'admission, ou ordonnance médicamenteuse d'admission (OMA). Le processus inclut (i) la décomposition des médicaments du BMO et de l'OMA en substances actives, (ii) la détection des substances actives arrêtées, ajoutées, modifiées et poursuivies entre le BMO et l'OMA, et (iii) l'analyse des divergences observées. EzMedRec a été évalué sur un échantillon de huit conciliations médicamenteuses incluant 74 prescriptions différentes avec un taux de conformité au gold standard défini par les pharmaciens de 98,64 %.

Mots-clés : modélisation de la prescription médicamenteuse, conciliation médicamenteuse, système d'aide à la décision, maladies chroniques, médicaments.

1 Introduction

La polypathologie est l'un des défis les plus importants des systèmes de santé actuels dans la gestion des maladies chroniques. Dans les pays développés, environ un adulte sur quatre souffre d'au moins deux maladies chroniques et plus de la moitié des personnes âgées souffrent d'au moins trois maladies chroniques (Hajat & Stein, 2018). Alors que la mortalité mondiale diminue, les gens vivent plus longtemps avec des handicaps et des comorbidités multiples, ce qui a d'importantes répercussions sur les besoins en soins au niveau mondial (Global Burden of Disease, 2018).

Les patients présentant plusieurs comorbidités sont engagés dans des parcours de soins complexes impliquant plusieurs établissements de santé, et en général au moins autant de professionnels de santé que les patients présentent de pathologies. Les divers fournisseurs de soins, par exemple, les hôpitaux et les cabinets médicaux, les laboratoires de biologie, les centres de radiologie, et les pharmacies utilisent habituellement différents systèmes de dossiers patient informatisés (DPI) pour reporter les éléments d'information relatifs à l'histoire de la maladie, l'examen clinique, les diagnostics et les traitements médicamenteux des patients. Bien que des progrès importants aient été réalisés, l'interopérabilité à l'échelle nationale des systèmes d'information de santé permettant l'échange d'informations appropriées dans un format numérique entre les différents acteurs de la prise en charge demeure complexe (Donahue et al., 2018). Les informations sont fragmentées et les professionnels de santé ne disposent souvent

que d'une partie des informations sur l'état d'un patient faisant de la coordination des soins pour les patients atteints de maladies multiples une tâche complexe. Du fait du manque de communication entre les différents acteurs de la prise en charge, les patients sont en effet exposés à une rupture dans la continuité des soins (Corbett et al., 2010). De nombreuses études ont en effet rapporté qu'il existait fréquemment des différences dans les prescriptions médicamenteuses aux points de transition de soins dont certaines étaient associées à des événements indésirables liés aux médicaments (Redmond et al., 2018).

La démarche de conciliation médicamenteuse a été proposée pour prévenir les erreurs médicamenteuses aux points de transition de soins. Des études ont démontré qu'un bon système de conciliation pourrait aider à réduire le taux d'erreurs médicamenteuses chez les patients hospitalisés de 76 % (Tam et al., 2005). Il s'agit notamment d'identifier et de résoudre les divergences non souhaitées entre les listes de médicaments à l'admission et à la sortie de l'hôpital. Les divergences comprennent, sans toutefois s'y limiter, les omissions, les ajouts, les redondances et les erreurs de dosage de médicaments. Si l'on considère le cas spécifique de l'admission à l'hôpital, la première étape du processus de conciliation médicamenteuse consiste à recueillir, en utilisant différentes sources d'information, la liste exhaustive et complète de tous les médicaments effectivement pris par le patient avant son hospitalisation. Cette liste est connue sous le nom de bilan médicamenteux optimisé ou BMO. L'élaboration du BMO repose habituellement sur l'interrogatoire du patient et sur l'analyse des ordonnances qu'il pourrait fournir. Il a été observé que cette étape de la conciliation médicamenteuse était souvent mal exécutée par les infirmières ou les médecins au moment de l'admission, ce qui crée un risque d'erreurs médicamenteuses dès l'admission à l'hôpital.

La conciliation médicamenteuse à l'admission à l'hôpital s'inscrit dans deux modèles différents (World Health Organization, 2014) : un processus proactif, lorsque le BMO est établi *avant* la rédaction de l'ordonnance des médicaments à l'admission (OMA), qui contribue à la prévention des erreurs médicamenteuses, et le modèle rétroactif, lorsque le BMO est établi *après* l'OMA, qui contribue à l'interception des erreurs médicamenteuses avérées. Dans le modèle rétroactif, le BMO est comparé à l'OMA pour identifier et résoudre les divergences non documentées, en faisant la différence entre les divergences non documentées *intentionnelles* pour lesquelles le prescripteur a délibérément fait le choix d'ajouter, de modifier ou d'arrêter un médicament, ce choix n'étant pas clairement documenté, et les divergences non documentées et *non intentionnelles*, lorsque le prescripteur change, ajoute ou arrête un médicament pris par le patient avant son admission alors qu'il pense continuer le traitement sans le modifier.

Il a été démontré que la conciliation médicamenteuse est une tâche pharmacologique chronophage et exigeante sur le plan cognitif, actuellement mal exécutée. Aussi, de nombreux outils ont été développés pour faciliter le bilan comparatif des prescriptions. Certains logiciels d'aide à la prescription ont d'abord proposé des extensions permettant d'ajouter des services d'aide à la conciliation médicamenteuse (Agrawal & Wu, 2009). Des logiciels spécialisés pour la conciliation médicamenteuse ont également été développés. Cependant, à la connaissance des auteurs, ces systèmes proposent essentiellement un pré-remplissage du BMO, par exemple, MedManage (Jarrett et al., 2019), l'outil de construction de la liste des médicaments à la pré-admission (Schnipper et al., 2009) ou RightRx (Tamblyn et al., 2018). De plus, la faible convivialité des outils existants a entravé leur adoption et leur efficacité (Boockvar et al., 2011) et des solutions plus récentes fournissant des interfaces ergonomiques ont été proposées (Plaisant et al., 2015 ; Horsky et al., 2018) qui laissent néanmoins à l'utilisateur la tâche lourdement cognitive d'identifier et de résoudre les divergences.

Afin d'alléger la lourdeur cognitive de cette tâche et d'améliorer la qualité de la conciliation médicamenteuse à l'admission à l'hôpital, nous avons développé un système d'aide à la conciliation médicamenteuse sémantiquement enrichie, le système EzMedRec appliqué à la conciliation médicamenteuse rétroactive. EzMedRec effectue une analyse formelle du BMO et de l'OMA afin d'identifier les divergences médicamenteuses, d'étiqueter les divergences identifiées selon une typologie que nous avons proposée et de fournir une résolution de ces divergences. En pratique, l'objectif est de créer la liste résultante des médicaments que le patient devrait prendre, à partir des deux listes de médicaments, le BMO et l'OMA. Cette liste finale

est a priori la liste des prescriptions adaptées sur le plan thérapeutique, exempte d'effets secondaires indésirables et d'interactions médicamenteuses. Elle doit être alors évaluée par le médecin pour identifier les divergences intentionnelles et valider la résolution proposée pour les divergences non intentionnelles.

2 Matériel et méthode

2.1 Matériel

Nous avons utilisé les documents pour la conciliation médicamenteuse produits par la Haute Autorité de Santé (HAS) (HIGH 5's Initiative, 2015) qui sont largement dérivés du « High 5s Standard Operating Protocol » de l'OMS (World Health Organization, 2014) où un formulaire destiné à être imprimé est utilisé comme support papier devant être rempli par les pharmaciens en charge de la conciliation médicamenteuse rétroactive à l'admission.

Le département de la pharmacie de l'hôpital Tenon à Paris nous a fourni un échantillon de huit formulaires de conciliation médicamenteuse rétroactive, dûment remplis, et anonymisés sur support papier. Cet échantillon a été utilisé comme gold standard pour la validation de notre outil EzMedRec. La figure 1 donne un exemple du formulaire utilisé à Tenon.

Nous avons utilisé trois ressources électroniques sur les médicaments : (i) la base de données publique française des médicaments développée par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) qui fournit la liste de tous les médicaments approuvés par la HAS (http://base-donnees-publique.medicaments.gouv.fr), (ii) la classification anatomique, thérapeutique et chimique (ATC) de l'OMS qui classe les médicaments en différents groupes selon l'organe, le système sur lequel ils agissent ou leurs caractéristiques thérapeutiques, pharmacologiques et chimiques, (iii) une table de correspondance entre les codes ATC et les noms commerciaux des médicaments.

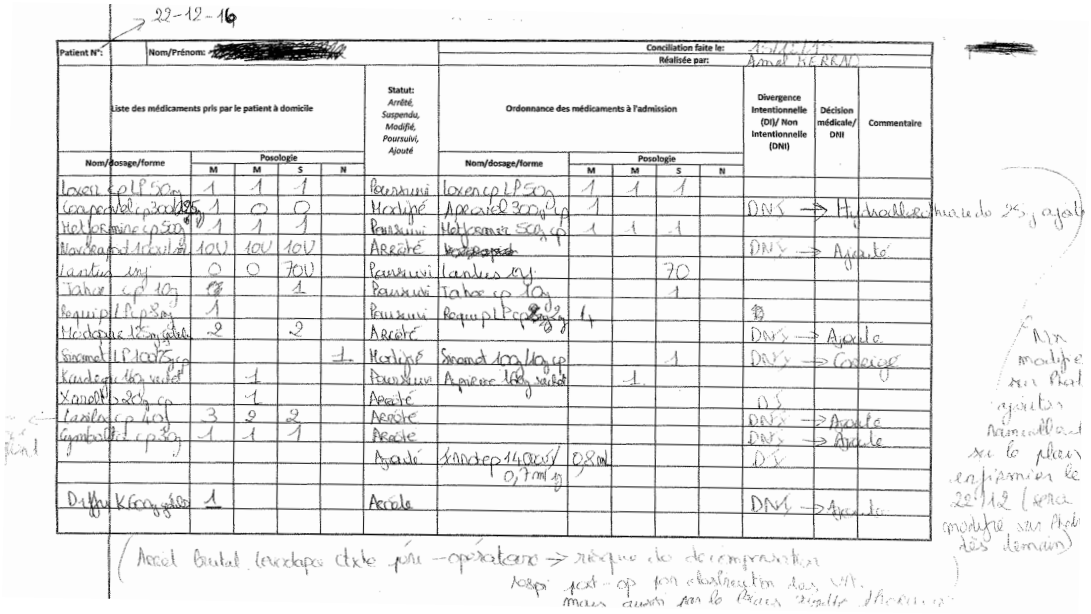


FIGURE 1 – Un exemple de formulaire HAS pour la conciliation médicamenteuse rétroactive à l'admission complété par les pharmaciens de l'hôpital Tenon (Paris, France).

2.2 Modélisation de la prescription

Afin de comparer les deux listes de médicaments, c'est-à-dire le BMO et l'OMA, nous avons utilisé un sous-ensemble des éléments constitutifs d'une ordonnance. En effet, si l'OMA est une liste de prescriptions, le BMO peut ne pas être constitué de prescriptions formelles, et apparaît souvent comme une liste de médicaments pris par le patient sans pour autant comporter tous les éléments d'information d'une prescription. Nous supposons que chaque médicament apparaissant dans le BMO et l'OMA contient les informations suivantes:

1 Le nom commercial du médicament, et ses substances actives, par exemple, COVERAM® (Périndopril arginine / Amlodipine).
2 La dose et l'unité du médicament et de chaque substance active, p. ex. COVERAM 10/5, 10 mg de Périndopril arginine, 5 mg d'Amlodipine.
3 La forme galénique (comprimés, gélules, etc.).
4 La voie d'administration (per os, intraveineux, etc.).
5 Le moment des prises et la quantité prescrite à chaque prise, par exemple le matin : 1 comprimé, le midi : 0, le soir : 1 comprimé, la nuit : 0.
6 La durée du traitement.
7 Des commentaires en langage naturel (optionnels) qui peuvent accompagner la prescription, par exemple prise du médicament en fonction de certaines conditions comme la fièvre, la douleur, etc.

2.3 Typologie des divergences

À partir d'une analyse de la littérature scientifique et des exemples de conciliation médicamenteuse fournis par la pharmacie de l'hôpital Tenon, nous avons proposé une typologie des divergences. La table 1 énumère 15 types de divergence entre le BMO et l'OMA qui peuvent être formellement détectées sur la base du modèle de prescription précédemment décrit. Les huit premiers types de divergences correspondent à ceux explicitement mentionnés dans le guide HAS de conciliation médicamenteuse. Les autres types de divergence sont des propositions originales et ont été ajoutés car ils correspondent à des situations qui peuvent potentiellement générer des erreurs médicamenteuses, en particulier en cas de médicaments combinant plusieurs substances actives qui peuvent être décomposés ou combinés (#12-15). Les différents types de divergences ne sont pas exclusifs, et on peut observer dans certains cas la co-occurrence de plusieurs divergences, par exemple un changement de posologie, de forme et de dose. De plus, certaines divergences formelles peuvent avoir un impact thérapeutique modeste. Par exemple, on peut observer un changement de rythme d'un même médicament, avec 1 pilule par jour changée en deux pilules par jour, mais la pilule prise une fois par jour est de 500 mg alors que chacune des deux pilules prises deux fois par jour est de 250 mg.

2.4 Principes de l'algorithme de conciliation médicamenteuse rétroactive enrichie

Le processus de conciliation médicamenteuse enrichie est formalisé sous la forme d'une séquence de cinq étapes. Les données entrantes sont le BMO et l'OMA, chacun étant structuré sous la forme d'une liste de prescriptions telles que précédemment décrites.

L'objectif de la conciliation médicamenteuse est de mettre en correspondance les lignes de prescription du BMO et celles de l'OMA. Avec EzMedRec, l'objectif est d'étendre le processus afin de traiter notamment le cas des décompositions / recompositions des médicaments combinant plusieurs substances actives (SA). Le résultat reste néanmoins le même avec des prescriptions étiquetées « arrêté », « ajouté », « poursuivi» et « modifié ». Une autre retombée du système consiste à produire une documentation des divergences observées, identifiées selon la typologie décrite dans la table 1 pour chaque ligne de prescription. Au cours du traitement, les comparaisons entre lignes de prescription sont conduites sur les médicaments décomposés selon leurs substances actives, mais les liens avec les prescriptions originales (médicaments) sont conservés.

TABLE 1 – Typologie des divergences pouvant être observées entre le BMO et l'OMA.

1	Ajout d'un nouveau médicament : un médicament absent du BMO est ajouté dans l'OMA.
2	Arrêt d'un médicament : un médicament présent dans le BMO est absent de l'OMA.
3	Modification de la forme galénique : le médicament est présent dans le BMO et l'OMA mais la forme galénique a été modifiée (par exemple, comprimés vs. gélules).
4	Modification de la voie d'administration : le médicament est présent dans le BMO et l'OMA mais la voie d'administration a été modifiée (par exemple, per os vs. intraveineux).
5	Modification de la dose : au moins une des substances actives a une dose différente dans le BMO et l'OMA (par exemple 500 mg vs. 250 mg).
6	Modification du rythme journalier des prises (par exemple 1 comprimé le matin vs. un comprimé matin et soir).
7	Modification de la dose à chaque prise (par exemple 500 mg le matin vs. 250 mg matin, midi et soir).
8	Modification de la dose journalière avec un surdosage ou un sous-dosage du médicament dans l'OMA par rapport au BMO (par exemple 2 g par jour vs. 3 g par jour).
9	Modification de la durée du traitement (par exemple prescription sur 1 semaine vs. 2 semaines).
10	Répétition d'un produit actif dans l'OMA : une substance active existe dans plusieurs médicaments de l'OMA (par exemple, le LOPRESSOR et l'AVALIDE contiennent tous les deux de l'hydrochlorothiazide).
11	Modification des commentaires : des instructions optionnelles sont différentes (par exemple « en cas de fièvre » vs. sans précision).
12	Arrêt d'une substance active. Par exemple, le médicament COAPROVEL 300mg/25mg (irbesartan 300 mg et hydrochlorothiazide 25 mg) du BMO reconduit en APROVEL 300 mg (irbesartan 300mg) dans l'OMA (l'hydrochlorothiazide 25mg serait arrêté).
13	Ajout d'une substance active. Par exemple, le médicament APROVEL 300mg (irbesartan) du BMO reconduit en COAPROVEL 300mg/25mg (irbesartan 300mg associé à l'hydrochlorothiazide 25mg) dans l'OMA (l'hydrochlorothiazide 25mg serait ajouté).
14	Changement d'une substance active par décomposition (en plusieurs médicaments). Par exemple, le médicament COAPROVEL 300mg/25mg (irbesartan 300 mg et hydrochlorothiazide 25 mg) du BMO reconduit en APROVEL 300 mg (irbesartan) et ESIDREX 25mg (hydrochlorothiazide) dans l'OMA.
15	Changement d'une substance active par re-composition (regroupement en un seul médicament). Par exemple, les médicaments APROVEL 300mg (irbesartan) et ESIDREX 25mg (hydrochlorothiazide) du BMO reconduits en COAPROVEL 300mg/25mg (irbesartan 300 mg et hydrochlorothiazide 25 mg) dans l'OMA.

Première étape – Prétraitement : Les SA des médicaments du BMO et de l'OMA sont extraites des prescriptions du BMO et de l'OMA. Chaque médicament D_i du BMO et de l'OMA est d'abord recherché dans la base de données publique des médicaments. Lorsque D_i est un médicament non combiné, Di est équivalent à une substance active unique, noté SA_i complétée par les données provenant de la base de données publique des médicaments telles que la dénomination commune internationale (DCI), la posologie et l'unité posologique de SA_i. Lorsque D_i est un médicament combiné, D_i est réécrit comme l'ensemble de toutes ses SA, notées SAi_j, et le même travail est effectué pour compléter chaque SAi_j. Dans ce dernier cas, le lien reliant chaque SAi_j au médicament original D_i est conservé. Chaque SA est également recherchée dans l'ATC afin d'y associer son code ATC. Le résultat du prétraitement est la production de deux listes de SA complétées et classées par ordre alphabétique, notées L_{BMO} et L_{OMA}.

Deuxième étape – Détection des ajouts, arrêts et répétitions de substances actives : Chaque SA de L_{BMO} est comparée aux SA de L_{OMA} afin d'identifier les arrêts de SA (lorsque la SA est présente dans L_{BMO} mais non présente dans L_{OMA}) et les ajouts de SA (lorsque la SA est présente

dans L_{OMA} mais non présente dans L_{BMO}). Les arrêts et les ajouts sont marqués pour être affichés comme "arrêté" ou "ajouté" lors de la spécification du type de divergence au niveau de l'interface utilisateur. La répétition d'une SA se produit lorsque la SA apparaît plus dans L_{OMA} que dans L_{BMO} sous la contrainte que la dose quotidienne cumulative de la SA dans L_{OMA} soit supérieure à la dose quotidienne cumulative de la même SA dans L_{BMO}.

Troisième étape – Spécification des prescriptions reconduites : Au cours de cette étape, nous considérons les SA qui ont été prescrites dans le BMO et l'OMA avec la même dose journalière cumulative. Elles sont étiquetées comme "poursuivi". Nous faisons la différence entre les prescriptions identiques et prescriptions similaires d'une SA. Les prescriptions identiques d'une SA partagent exactement les mêmes modalités de prescription en termes de dose, fréquence, rythme d'administration, forme, etc. Lorsqu'au moins un critère de la prescription d'une SA est différent, les prescriptions sont considérées comme similaires. Les prescriptions identiques et les prescriptions similaires qui sont poursuivies sont affichées dans des cases de couleur. Les divergences existant entre des prescriptions similaires sont considérées comme mineures et sont signalées en tant que telle au niveau de l'interface utilisateur.

Quatrième étape – Détection des modifications de prescriptions : Les changements de prescriptions sont soulignés en mettant l'accent sur les divergences affectant les mêmes SA avec des doses journalières cumulatives différentes. Ces prescriptions sont affichées dans un espace différent afin de rendre visible très rapidement les sous-dosages et les surdosages. Tout comme pour les médicaments poursuivis, les médicaments qui contiennent les mêmes SA sont regroupés. Au cours de cette étape, dès qu'un médicament combiné est présent, on mettra en œuvre la recomposition.

Cinquième étape – Détection des prescriptions médicamenteuses équivalentes sur le plan thérapeutique : Les prescriptions arrêtées ou ajoutées sont analysées afin de déterminer si les divergences observées peuvent s'expliquer comme des substitutions de médicaments servant le même objectif thérapeutique. Pour chaque SA de L_{BMO} arrêté dans la L_{OMA} et pour chaque SA ajouté dans la L_{OMA}, un appariement des codes ATC est mis en œuvre à différents niveaux d'abstraction. Cet appariement sémantiquement contraint repose sur la structuration de l'ATC en fonction des propriétés thérapeutiques des SA. Lorsque l'appariement est réalisé à un faible niveau d'abstraction, on considère que la divergence observée pourrait être intentionnelle et correspondre à une substitution. En pratique, nous avons considéré que le troisième niveau (sous-groupe pharmacologique) de la structure ATC était le niveau le plus élevé qui pouvait être conservé pour proposer l'hypothèse d'une substitution.

3 Résultats

3.1 Le système EzMedRec

Le système EzMedRec offre plusieurs affichages du BMO et de l'OMA, ainsi que les résultats de la conciliation médicamenteuse en fonction de l'étape du processus. L'affichage le plus significatif est similaire au formulaire de conciliation médicamenteuse promu par la HAS avec une présentation tabulaire. Les principales colonnes sont le BMO, l'OMA et les divergences identifiées. Les lignes du BMO et de l'OMA correspondent à des médicaments (composés d'une seule SA ou d'une combinaison de plusieurs SA) avec les mises en correspondance entre le BMO et l'OMA lorsque cela est possible. Les figures 2, 3 et 4 donnent un exemple du processus de conciliation médicamenteuse rétroactive enrichie pour un BMO composé de cinq prescriptions médicamenteuses et l'OMA associée, composée de six prescriptions médicamenteuses.

Dans un premier temps, le BMO et l'OMA sont affichés sous la forme de listes de prescriptions (voir figure 2). Le nom du médicament est donné avec la forme, la voie d'administration, le rythme, la dose, et la durée du traitement.

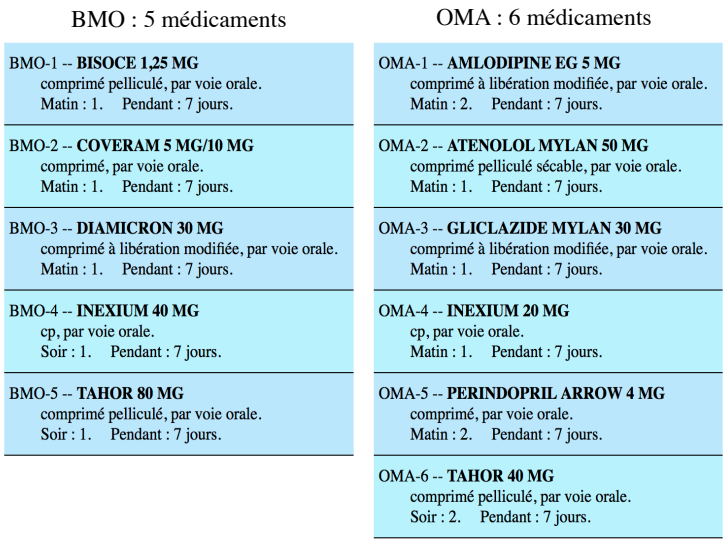


FIGURE 2 – Un exemple de BMO avec son OMA avant la conciliation médicamenteuse.

Puis, le système affiche les résultats du processus de conciliation médicamenteuse, en mettant en évidence les divergences marquées comme « arrêté», « ajouté » et « modifié ». Les autres lignes de prescriptions sont étiquetées "poursuivi". La figure 3 présente les résultats les résultats de la conciliation médicamenteuse pour le BMO et l'OMA présentés à la figure 2. Les codes de couleur correspondent aux types de transition entre les prescriptions du BMO et de l'OMA, rouge représente les médicaments arrêtés, vert les médicaments ajoutés, orange les médicaments modifiés, et bleu les médicaments poursuivis (identiques ou similaires).

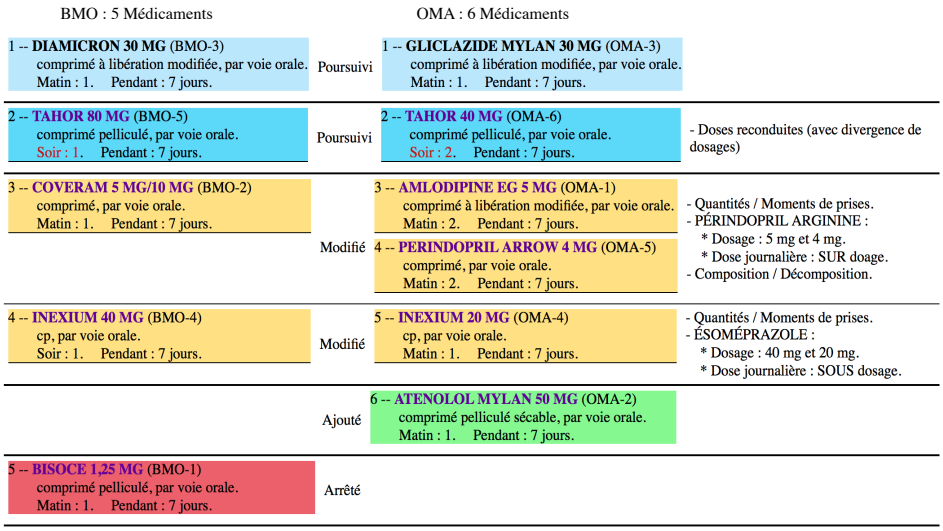


FIGURE 3 – Exemple d'un BMO et d'une OMA après conciliation médicamenteuse : affichage des divergences.

3.2 Détection des substitutions potentielles

On peut émettre l'hypothèse d'une substitution potentielle lorsqu'un médicament arrêté dans le BMO et un médicament ajouté dans l'OMA appartiennent à une même classe thérapeutique dans la hiérarchie ATC. Si l'on considère l'exemple de la Figure 3, on peut voir que BISOCE

(BMO-1) est arrêté pendant que l'ATENOLOLOL MYLAN (OMA-2) est ajouté. La SA du premier est le bisoprolol, la SA du second est l'aténolol. Le bisoprolol et l'aténolol sont tous les deux des bêta-bloquants sélectifs, avec une classe ATC mère commune (C07AB). Par conséquent, cette situation peut être identifiée comme une substitution potentielle de médicaments dans un même but thérapeutique. Comme le montre la figure 4, cette substitution potentielle est présentée à l'utilisateur pour validation.

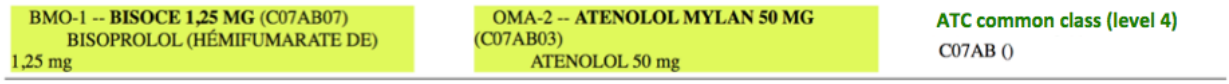


FIGURE 4 – Exemple d'un couple de divergences « arrêté » / « ajouté » considéré comme une substitution médicamenteuse. Le niveau d'équivalence de l'appariement est indiqué dans l'interface.

3.3 Évaluation préliminaire

Nous avons utilisé l'échantillon de 8 conciliations fourni par les pharmaciens de l'hôpital Tenon. Il est constitué d'un total de 74 lignes de prescription, 41 dans les BMO et 33 dans les OMA. La table 2 détaille les analyses effectuées par EzMedRec. Sept conciliations étaient strictement identiques à celles des pharmaciens. Une ligne de prescription, parmi les 74, a été mal classée, conduisant à un taux de classification correcte de 98,64 %. Cette ligne, classée « arrêtée » dans le gold standard, a été classée « modifiée » par EzMedRec. Dans cas précis, le BMO contenait 2 médicaments, SINEMET et MODOPAR, avec l'un « poursuivi » dans l'OMA et l'autre « arrêté ». Cependant, ces deux médicaments contiennent une SA identique, le levodopa. La quantité globale de levodopa était réduite dans l'OMA, mais sans que cela soit mentionné explicitement dans la conciliation qui a servi de gold standard. EzMedRec a effectué le regroupement du SINEMET et du MODOPAR dans le BMO et son association avec le SINEMET seul dans l'OMA avec le statut « modifié » en notant le sous-dosage de levodopa.

TABLE 2 – Résultats obtenus avec EzMedRec en comparaison au gold standard.

	Gold Standard		***EzMedRec***	
	BMO	**OMA**	**BMO**	**OMA**
Médicaments poursuivis	20	20	20	20
Médicaments modifiés	**6**	7	**7**	7
Médicaments arrêtés	**15**	–	**14**	–
Médicaments ajoutés	–	6	–	6
Nombre total de lignes de prescription	41	33	41	33
SA arrêtées	N/A	–	2	–
SA ajoutées	–	N/A	–	1

4 Discussion et Conclusion

Dans le contexte de la conciliation médicamenteuse rétroactive à l'admission, nous avons proposé une typologie des divergences formelles pouvant être détectées entre deux listes de médicaments, le BMO et l'OMA. La reconnaissance automatique de ces divergences a été implémentée dans l'outil EzMedRec selon les trois catégories habituelles : arrêté, ajouté et modifié, la quatrième étiquette résultant de l'absence de divergence étant notée « poursuivi ». Toutefois, la prise en compte des SA a permis une analyse plus fine qui tient compte des décompositions et recompositions des médicaments combinés. De plus, l'utilisation de l'ATC comme ressource sémantique basée sur la subsomption des classes thérapeutiques permet à l'outil de suggérer des substitution potentielles de SA lorsqu'une SA était arrêtée dans le BMO et qu'une autre SA était ajoutée dans l'OMA avec un même effet thérapeutique.

Au niveau de l'interface, les regroupements et l'affichage du détail des divergences détectées permettent de justifier le choix de la catégorie. Toutefois, tout comme les autres outils, EzMedRec ne permet pas d'identifier le caractère intentionnel ou non d'une différence et, à ce titre, nécessite la validation d'un médecin ou d'un pharmacien. EzMedRec utilise les ressources médicamenteuses de l'ANSM et par conséquent dépend de la structure et du contenu de ces bases. Si l'information sur le médicament est partagée au niveau mondial, en revanche, les modèles de données sur le médicament des bases de référence utilisées dans les pays sont différents. Par exemple, RxNorm aux États-Unis fournit un modèle structuré du médicament qui peut être utilisé directement par les systèmes informatisés. Dans l'état actuel, EzMedRed n'est pas adapté à RxNorm. L'alignement des ressources nationales sur un modèle commun et partagé permettrait le développement d'outils génériques pour aider les processus liés à l'usage des médicaments, comme la conciliation médicamenteuse.

La petite taille de l'échantillon utilisé pour valider EzMedRec est une limite de ce travail et une évaluation sur un nombre plus conséquent serait souhaitable. Cependant, un frein à l'obtention d'un échantillon plus large, au moins à l'hôpital Tenon, est la non disponibilité de BMO informatisés. Afin de combattre la iatrogénie médicamenteuse, la France a développé au niveau national le dossier pharmaceutique (DP) pour enregistrer toutes les prescriptions délivrées au niveau des pharmacies d'officine. Cependant, certains hôpitaux rencontrent des difficultés pour s'y connecter. C'est le cas à l'hôpital Tenon où l'information sur l'historique médicamenteux n'est pas récupérée depuis des ressources électroniques. Actuellement, EzMedRec est un prototype. Une évaluation par des professionnels de santé impliqués dans la conciliation médicamenteuse serait nécessaire pour valider l'exactitude des résultats et l'utilité d'un tel système, en particulier sur les aspects innovants comme la suggestion des substitutions.

Ce travail a été réalisé pour la conciliation rétroactive en se focalisant sur l'identification des différences entre BMO et OMA. Une perspective serait d'utiliser l'outil dynamiquement pour fournir une aide à la décision pour la conciliation proactive. Le BMO pourrait être utilisé comme point de départ de l'OMA à construire et chaque modification du BMO pourrait être détectée et qualifiée afin d'informer l'utilisateur sur les effets de ces modifications du point de vue de la conciliation.

Remerciements

Nous remercions le Docteur Amel Kerrad, pharmacienne de l'hôpital Tenon pour nous avoir fourni et commenté l'échantillon des huit conciliations médicamenteuses.

Références

AGRAWAL A, WU WY. Reducing medication errors and improving systems reliability using an electronic medication reconciliation system. *Jt Comm J Qual Patient Saf*. 2009;35(2):106–14.

BOOCKVAR KS, SANTOS SL, KUSHNIRUK A, et al. Medication reconciliation: barriers and facilitators from the perspectives of resident physicians and pharmacists. *J. Hosp. Med.,* 6 (6) (2011), pp. 329-337.

CORBETT CF, SETTER SM, DARATHA KB, NEUMILLER JJ, WOOD LD. Nurse identified hospital to home medication discrepancies: implications for improving transitional care. *Geriatr Nur (Lond)* 2010;31(3):188–96.

DONAHUE M, BOUHADDOU O, HSING N, TURNER T, CRANDALL G, NELSON J, NEBEKER J. Veterans Health Information Exchange: Successes and Challenges of Nationwide Interoperability. *AMIA Annu Symp Proc*. 2018 Dec 5;2018:385–94. eCollection 2018.

Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015;386(9995):743.

HAJAT C, STEIN E. The global burden of multiple chronic conditions: A narrative review. *Prev Med Rep*. 2018 Dec; 12: 284–93.

HIGH 5s Initiative Medication Reconciliation, Sept. 2015, https://www.has-sante.fr/portail/upload/docs/application/pdf/2016-04/traduction_rapport_medrec.pdf [Accessed March 11th 2019].

HORSKY J, DRUCKER EA, RAMELSON HZ. Higher accuracy of complex medication reconciliation through improved design of electronic tools. *J Am Med Inform Assoc*. 2018 May 1;25(5):465-475.

JARRETT T, COCHRAN J, BAUS A, DELMAR K. MedManage: The development of a tool to assist medication reconciliation in a rural primary care clinic. *J Am Assoc Nurse Pract*. 2019 Feb 27.

PLAISANT C, WU J, HETTINGER AZ, POWSNER S, SHNEIDERMAN B. Novel user interface design for medication reconciliation: an evaluation of Twinlist. *J Am Med Inform Assoc*. 2015 Mar;22(2):340-9.

REDMOND P, GRIMES TC, MCDONNELL R, BOLAND F, HUGHES C, FAHEY T. Impact of medication reconciliation for improving transitions of care. *Cochrane Database Syst Rev*. 2018 Aug 23;8:CD010791. doi: 10.1002/14651858.CD010791.pub2.

SCHNIPPER JL, HAMANN C, NDUMELE CD, LIANG CL, CARTY MG, KARSON AS, et al. Effect of an electronic medication reconciliation application and process redesign on potential adverse drug events: a cluster-randomized trial. *Arch Intern Med*. 2009 Apr 27; 169(8):771-80.

TAM VC, KNOWLES SR, CORNISH PL, FINE N, MARCHESANO R, ETCHELLS EE. Frequency, type and clinical importance of medication history errors at admission to hospital: a systematic review. *CMAJ*. 2005;173(5):510–5.

TAMBLYN R, WINSLADE N, LEE TC, MOTULSKY A, MEGUERDITCHIAN A, BUSTILLO M, et al. Improving patient safety and efficiency of medication reconciliation through the development and adoption of a computer-assisted tool with automated electronic integration of population-based community drug data: the RightRx project. *J Am Med Inform Assoc*. 2018 May; 25(5): 482–495.

World Health Organization. The High5s Project Standard Operating Protocol Assuring Medication Accuracy at Transitions in Care. 2014 Sept. http://www.who.int/patientsafety/implementation/solutions/high5s/h5s-sop.pdf. [accédé le 13 mai 2019].

Temporal models of care sequences for the exploration of medico-administrative data

Johanne Bakalara[1,2], Thomas Guyet[1], Olivier Dameron[1]
Emmanuel Oger[2,3], André Happe[2,4]

[1] UNIV RENNES, INRIA, CNRS, IRISA
[2] UNIV RENNES, EA-7449 REPERES
[3] CHU RENNES
[4] CHRU BREST
johanne.bakalara@irisa.fr

Abstract : Pharmacoepidemiology with medico-administrative databases enables to study impact of health products in real-life setting. These studies require to manipulate the raw data, the care trajectories, in order to identify pieces of data that may witness the medical information that is looked for. The manipulation can be seen as a querying process in which a query is a description of a medical pattern (*e.g.* occurrence of illness) with the available raw features from care trajectories (*e.g.* occurrence of medical procedures, drug deliveries, etc.). The more expressive is the querying process, the more accurate is the medical pattern search. The temporal dimension of care trajectories is a potential information that may improve the description of medical patterns. The objective of this work is to propose a formal framework that would design a well-founded tool for querying care trajectories with *temporal medical patterns*. In this preliminary work, we present the problematic and we introduce a use case which illustrates the comparison of several querying formalisms.

Mots-clés : Temporal logics, Description logic, Ontologies, Pharmacoepidemiology, SNDS.

1 Introduction

Pharmacoepidemiology studies uses and effects of drugs on population in real life including benefits and risks. As such, pharmacoepidemiology deals with positive impact (*i.e.* prevention of disease) as well as safety concerns. Patients with health events are identified, according to their individual characteristics, their treatments and their concomitant treatments. Collecting information to answer one epidemiological question requires a lot of time and is very expensive.

Medico-administrative databases are a potential alternative which are attractive because of their large population coverage and their availability. For instance, the SNDS[1] (Tuppin *et al.*, 2017) database contains individual information of French patients: age, sex, location; and health reimbursement information: drug deliveries, medical acts or medical visits and hospitalisations (date of arrival, leaving date, diagnosis code) but it does not contain medical reports.

The interest of using this medico-administrative database has been demonstrated by the suspension of benfluorex (Weill *et al.*, 2010). It was the first large-scale pharmacoepidemiology study in France that was possible thanks to information contained in the SNDS.

The study shows that diabetic patients exposed to Benfluorex have an higher risk of hospitalisation for heart disease than the unexposed diabetics in the following years. Epidemiologists use the SNDS database to find patients that experienced some medical events of interest. It is worth noticing that information has been collected for administrative purposes (care reimbursements), but not for medical ones. As a consequence, the medical content associated with medical events is relatively poor.

[1]SNDS: French National System for Health Data (previously SNIIRAM). The SNDS is the world largest medico-administrative database with a population coverage close to 99%.

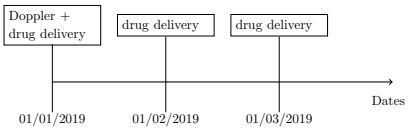


Figure 1: Example of a care trajectory.

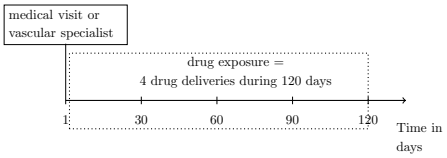


Figure 2: Example of a care pathway.

Let us consider a patient having a Venous Thromboembolism (VTE). The database does not contain the information that *the patient has the disease VTE*. This information has to be deduced from the information about its care deliveries. For instance, a consultation with a vascular specialist and deliveries of anticoagulant drugs. These pieces of information are available in the data in a structured format. Compared to medical record in hospital, the SNDS database has less medical information, but it does not required sensitive text analysis.

All medical events that are stored in the database compose the care trajectory of a patient. Figure 1 illustrates a care trajectory. The challenge of the epidemiologist is to define selection criteria that would reconcile those actual patient information with the medical semantic. The definition of these criteria composes a health pattern called care pathway illustrated Figure 2. A possible care pathway of patients with VTE[2] is a consultation with a **vascular specialist** or at the hospital with an **Doppler imagery act** (CCAM[3] *EDQM001* code) followed by at least 3 deliveries of **anticoagulant drugs within 4 months** (with the first delivery of anticoagulant **in the week after** consultation/Doppler). The care trajectory denotes the actual medical events stored in the database while a care pathway is a pattern of medical events that describe the pathology profile as illustrated on Figures respectively 1 and 2.

We aim at enabling epidemiologists to query the database of care trajectories with such kind of complex descriptions of care pathway. The complexity is twofold:

- use of ontological concepts (Doppler imagery act/anticoagulant/vascular specialist): the code of the medical act is given, but the code for anticoagulant drugs is not precisely given. Anticoagulants refer to a class of drugs that is described in the ATC taxonomy.[4]
- use of temporal constraints (in the week after/within 4 months): the temporal order of cares, numerical duration/delays specifies the temporal organisation of the events.

Our work focuses on having expressive temporal constraints in the description of care pathways. Temporal constraints enable to discriminate care trajectories of interest from care trajectories with same events but presenting different delays (*e.g.* an Doppler imagery act followed by an anticoagulant delivered several weeks later does not witness a VTE). This may help to specify care pathways that match only the desired care trajectories. It is worth noting that drugs deliveries, medical acts and hospital stays are available in the database with timestamps.

The challenge we face is summed up in Figure 3. Care pathways are on the left side and care trajectories are on the right side. The overall objective is to bridge the two sides. On the top in red, epidemiologists formulates a medical hypothesis. In blue, at the second level, epidemiologist describes care pathways with the available health information and a database stores health information. In green, at the third level, data and queries are specified in a formalism. This latter has to express a maximum of information to represent the complexity of the data (care trajectories) and of the care pathway. Green and blue levels are intertwined. Today, the tools for querying SNDS (third level) lack of expressiveness and constraint epidemiologists for their specification (SQL, SAS, etc...).

[2]This description is an illustration for pedagogical purposes, we refer to the use case for a medical description of VTE.

[3]CCAM: Classification Commune des Actes Médicaux

[4]ATC: Anatomical Therapeutic Chemical Classification System.

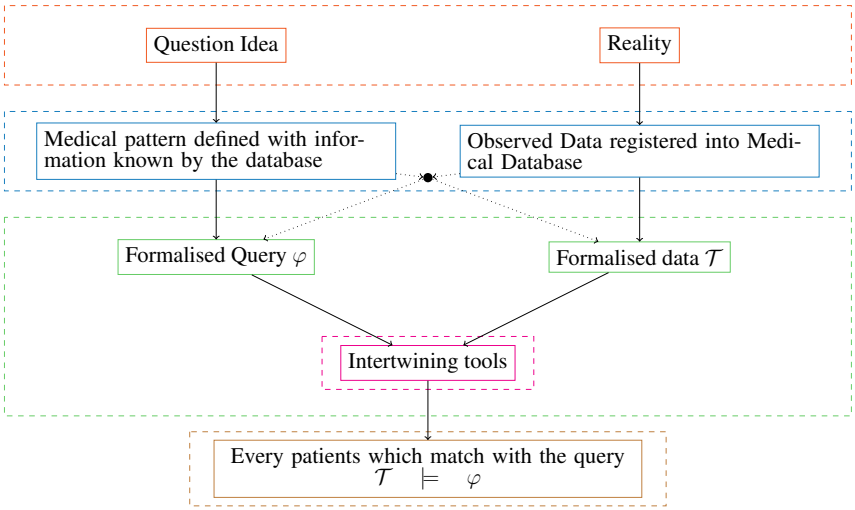


Figure 3: From medical study issue to the care trajectory querying

In this article, we show that the problem of querying patient care trajectories with temporal medical patterns (the care pathways) may be addressed, sometimes only partially, by tools or formalisms coming from different computer science fields: complex event processing, formal reasoning, knowledge representation and database. In this preliminary work, our objective is to identify strengths and weaknesses of these approaches.

In the next sections, we states formally of problem of querying a set of care pathways, then Section 3 reviews related work. The last section illustrates some of approaches of the related work on a case study.

2 Problem statement

The objective of this work is to propose a formal framework that would design a well-founded and efficient tool for querying care trajectories in the context of pharmacoepidemiology.

Generally speaking, let $\mathcal{T} = (T_i)_{i \in [n]}$ be a set of n care trajectories and φ a care pathway abstract description. φ holds in a care trajectory $T \in \mathcal{T}$, denoted $T \models \varphi$, if and only if the care trajectory *contains* the care pathway. The formalisation problem is threefold:

- define a formalism to model care trajectories, T, which represent the SDNS data
- define a formalism to model care pathways, φ, which specifies an abstract care pathways
- define a computational model that can evaluate whether T entails φ: $T \models \varphi$.

As we noticed in the introduction, specifying care pathways requires to manipulate: temporal concepts (time constraints and time window), medical concepts and knowledge (ontologies). The ideal formal framework should capture these dimensions, enable intuitive queries to be expressed for a wide range of pharmacoepidemiological studies and be computationally efficient.

It is of paramount importance to base choices on solid theoretical foundations. Expressiveness and efficiency are known to be antagonist objectives (Levesque, 1986). A well-founded approach would be the basis for proposing long-term solutions, make possible future improvements and facilitate its application to a broad range of contexts (*i.e.*, various databases, queries).

3 Related work

This section presents four families of formalisms to answer the problem: model checking; Complex Event Processing (CEP); temporal databases; and Knowledge Representation and Reasoning (KR). The work that we mentioned addresses the three problems mentioned in Section 2 at the same time. The formalism should represent data (care trajectories), query (care pathway) and to compute the answers of the queries on data. The last two families have been more explored in medical context (Combi *et al.*, 2010) than the two others.

Model checking (Clarke Jr *et al.*, 2018) verifies if a model satisfies a property or a formula. This research line is interested in representing dynamic systems with formal temporal formalisms (discrete event models, $\mathcal{M}$, describing how the system evolves) to prove some properties specified by formula φ. A formula φ is true if and only if φ is true for any traces [5] that can be generated from the model $\mathcal{M}$. The most common formalisms for formula in Model Checking are LTL (Linear Temporal Logic), CTL (computation tree logic) or MTL (metric temporal lgic) which is the temporal extension of LTL.

To apply such methods in our context, the events of care trajectory are represented by one finite trace of the system (and there is no system model in our case) and the care pathways is represented by a temporal logic formula. The care trajectory is selected if and only if the trace satisfies the formula. These methods are interesting because they provide formal results (expressiveness, completeness, equivalences) on the representation of timed systems, but they don't manage neither reasoning nor ontological representation. In related medical domain, model checking has been used to study the compliance of care pathways (Bottrighi *et al.*, 2010).

Complex Event Processing (CEP) (Giatrakos *et al.*, 2017) is a research line that aims at processing log-streams with patterns. Log-streams are streams or sequences of timed events. The CEP processes these logs to detect or to locate complex events (or *patterns*) defined by the user. This domain defines formalisms that aim at being very efficient to process streams and expressive to specify patterns. Temporal constraint networks (Cabalar *et al.*, 2000) or Chronicles (Dousson & Le Maigat, 2007) are simple temporal models that are interesting for their graphical representation, but are limited to simple relational events. Some more complex formalisms, *e.g.* ETALIS (Anicic *et al.*, 2011) or logic-based event recognition (Giatrakos *et al.*, 2017), propose very expressive representations of complex events, including reasoning techniques (including ontologies) which enrich the capabilities of CEP.

In our context, care trajectories are logs, and care pathways are the complex events. We are not interested in the stream dimension of these approaches, but their formalisms to represent complex events may be adapted in the context of static logs.

Temporal databases (Snodgrass, 1992) extend the notion of database to timestamped data. Databases issued data representation problems but also specific querying language problems. We gather in this family the temporal extension of relational databases (*e.g.* TSQL) but also web semantic approaches which combine query languages (*e.g.* SPARQL) and expressive description languages. Care trajectories are facts in the temporal database and the querying of care pathways becomes a problem of specifying care pathways in the query language. Rivault *et al.* (2019) shown that semantic web is an interesting approach for our problem, but does not explicitly address the problem of timed queries.

Finally, **Knowledge Representation and Reasoning** (KR) (Levesque, 1986) is "the study of how what we know can at the same time be represented as comprehensibly as possible and reasoned with as effectively as possibly". In this research domain, temporal KR is focused on representing and reasoning about time. It gives rise to several logics (Long, 1989), for instance: Allen's logic, McDermott's logic, Event Calculus or Halpern & Shoham's logic.

[5] Sequences taht register the set of atomic propositions that are valid along the execution, *cf.* part 3.2.2 in (Baier & Katoen, 2008)

KR is a general framework to study how to represent care trajectories and how to model reasoning-based queries on care pathways. Approaches from the other families may be represented with appropriate logics. Studying KR formalisms seems of paramount importance as it provides common foundations to compare various approaches. Description Logic (DL) (Baader *et al.*, 2003) is a KR formalism allowing ontology-mediated query answering (OMQA) (Bienvenu, 2016). Artale et *al.* (Artale *et al.*, 2017) present a temporal extension of DL that may be suitable for our problem. For instance, (O'Connor *et al.*, 2009) developed a tool based on OWL for research data management with a temporal reasoning in a clinical trial system.

4 Comparison of approach on a use case

4.1 Rational

This section introduces a real use case. In this example, pharmacoepidemiologists want to select patients with Venous Thromboembolism (VTE) from the data contained in the SNDS. Venous thromboembolism, *i.e.* deep vein thrombosis (DVT) or pulmonary embolism (PE), is a frequent and potentially fatal disease (Oger & EPI-GETBO, 2000; Delluc *et al.*, 2018). This requires to survey how many people are concerned, if the number of patients increased and if a specific drug has an impact. The difficulty for epidemiologists lies in the description of the care pathways that will accurately identify VTE from the SNDS data. The description below describes two care pathways that physicians proposed to identify VTE (referring to Figure 3, this description is between the red and blue levels).

> *In clinical practice facing a clinical suspicion of VTE, physicians first prescribe anticoagulant and then confirm or not the diagnosis through specific medical acts: for instance Doppler ultrasonography or CT scan. Patients with suspected PE are often hospitalized whereas patients with suspected DVT are managed on a ambulatory basis. If the suspicion is confirmed, anticoagulant deliveries continues for 3 to 12 months or sometimes longer duration. Hence, diagnosis (through medical act) is preceded or followed by anticoagulant initiation within a time window of at most 0 to 2 days, keeping in mind that PE suspicion leads to hospitalisation during which medical act to confirm the diagnosis are performed and then anticoagulant is observed only after the patient comes back home.*

Through these observations, pharmacoepidemiologists identified the following two care pathways to detect patients with VTE form SNDS data (referring to Figure 3 this description is in the blue level):

1. A diagnosis (DVT or PE) or a medical act (Doppler or CT scan) during or prior to anticoagulant(AC) deliveries for 1 to 2 days and delivery lasts a minimum of 3 months and a maximum of 12 months (sometimes longer). Each delivery is separated by 0 to 2 months.

2. A diagnosis PE during an hospitalisation followed by AC delivery.

These care pathways contain sequential order but also time constraints between events (for instance number of days) or duration of events (time window). Searching for such patterns requires high expressiveness that make databases query languages (SQL, TSQL) practically difficult to use. Next section illustrates alternative solutions to represent the green level (see Figure 3).

This use case illustrated the problem of formalizing care pathways of patients suffering from VTE. But, for sake of generality, our formalism has to specify the care pathway patterns of as many case studies as possible.

4.2 Comparison of different models

This section aims at illustrating the comparison of four formalisms, Description Logic (DL), Chronicles, LTL and MTL to discuss about their power of expressivity in our case of study.

DL supplies ontologies query answering technologies for Ontology-Based Data Access (OBDA) and is well known to represent medical data. Indeed, Description Logic is used to describe and reason about concepts on data. Reasoning with Description Logic is performed in three steps. The first one consists in defining the data and the data form (called *ABox*) which contains knowledge at the instance level: a set of assertion defining concepts, roles and a countably infinite set of individuals names. Concepts with individuals names and roles with individuals names are forming atoms.

The second step consists in define a base of knowledge (*TBox*) which is a set of concepts inclusions. Concept inclusions represent a hierarchy of concepts. For instance, the concept $B01AF01$ representing anticoagulant drugs is the leaf concept in the hierarchy of concept modeling the ATC classification:

- B: Blood and blood forming organs
  - B01: Antithrombotic agents
    - B01AF: Direct Xa inhibitor
      - B01AF01: Rivaroxaban

And considering the CCAM (medical acts) code for the Doppler: EDQM001 (iliac and lower limb arteries) we could construct the following *ABox*, where Pierre and Paul are patients and n_i are dates of medical events:

$$B01(\mathsf{Patient1}, t_1) \quad B01(\mathsf{Patient1}, t_2) \quad PE(\mathsf{Patient1}, t_3)$$
$$B01AF01(\mathsf{Patient2}, t_1) \quad B01AF01(\mathsf{Patient2}, t_3) \quad EDQM001(\mathsf{Patient2}, t_4)$$
$$t_1 < t_2 < t_3 < t_4$$

And *TBox* issued from the previous ATC classification:

$$B01AF01 \sqsubseteq B01AF \quad B01AF \sqsubseteq B01 \quad B01 \sqsubseteq B$$

$B01AF01$ is the subclass of $B01AF$ which is the subclass of $B01$ which is the subclass of B. Operators linking concepts are defined depending on the class of DL chosen. For example, the $\mathcal{ALC}$ family offers the concept constructors: negation, conjunction, disjunction, existential restriction $\exists$ and universal restriction $\forall$.

The third step is to define a query to extract information from knowledge contained in *ABox* and *TBox*. Usually, queries are expressed with a first order logic. Artale *et al.* (2017) designed a temporal DL: TQL that extends the standard ontology language: OWL 2QL. It offers the capability of having time as individuals names and to compare whether an atom occurs before another. In our case study, we propose a query to find patients designed by the use case defined in Section 4.1.

$$\exists patientID, t_1(t_{ref} < t_1) \wedge EDQM001(patientID, t_{ref}) \wedge B01(patientID, t_1)$$
$$\wedge t_2(t_2 < t_1) \wedge B01(patientID, t_2)$$
$$\wedge t_3(t_3 < t_1) \wedge B01(patientID, t_3)$$

We find patients having three deliveries of $B01$ (and subclasses of $B01$) in a row and the first one is delivered after an Doppler. The temporal information that can be represented here is limited to the order of the events. Delay or duration can not be specified.

To express queries with temporal delays, the formalism of chronicle represents a care pathway as a temporal constraint graph. Chronicles allow the expression of sequential order of events with temporal constraints such as interval of time. Furthermore, negative time in the

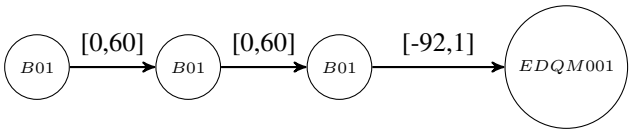


Figure 4: Chronicles

interval expresses that an event may occur before or after another one. Figure 4 specifies patients having at least three anticoagulant deliveries separated by 0 to 60 days, and a diagnosis DVT before, after or during deliveries. DVT occurs 92 days earlier or one day after the third delivery.

However, we can not explicitly restrict the number of deliveries to 12 months as defined in the use case. We also cannot use the notion of *no event* (event does not occur). Model checking offer the possibility to express *no event* and can be used as queries. Such as an example we propose the following LTL formula as an example applied to our case of study:

$$(\Diamond D_{B01} \wedge \bigcirc(\Diamond D_{B01}) \wedge \bigcirc(\Diamond D_{B01}) \wedge (\Diamond(DVT \vee PE))$$

The LTL formula represents a care pathway with at least three deliveries and a diagnosis DVT or PE. We literally read it: *in the future ($\Diamond$), there is the delivery of $B01$ and ($\wedge$) it is followed ($\bigcirc$), in the future, by the delivery of $B01$ and it is followed, in the future, by a delivery of $B01$ and, in the future, there are the diagnosis DVT or ($\vee$) PE)*. LTL only contains order between events and doesn't contains time constraints. It is quite limited for our problem, so we refer to its temporal extension MTL. The MTL formula adds the capability to express quantitative temporal constraints. We propose the following MTL formula as an example applied to our case of study:

$$\Diamond(DVT \vee PE) \Rightarrow ((\Diamond_{[0\ 2]} D_{BO1}) \wedge (\Diamond_{[0\ 60]} D_{BO1}) \wedge (\Diamond_{[0\ 60]} D_{BO1}) \wedge (\Diamond_{\geq 365}\ (\Box(\neg D_{BO1}))))$$

It represents a care pathway with a DVT or PE followed between 0 to 2 days after by three AC deliveries separated between 0 to 60 days, and no deliveries occur after 365 days. MTL can explicitly restrict the number of deliveries and temporal constraints but the notion of sequences is manually found by the multiple use of $\Diamond$. For deep understanding of notations, we refer the reader to Bouyer *et al.* (2005).

From computational point of view, Chronicles may be very space/time-efficiently to be recognizes in care trajectories. Simple LTL formula would also be space/time-efficient to check but it is expressively poor. In contrast, MTL is known to be undecidable. It is a theoretical limitation but, not necessary a practical constraint Ouaknine & Worrell (2005).

5 Conclusion

In this article, we introduced the context of pharmaco-epidemiological studies with medico-administrative databases and the challenge to query such databases with medical questions. It consists in finding patients satisfying a medical pattern that we want to be expressive.

To express medical patterns, we use the formalism of Description Logic to describe data and include medical knowledge issued from the available classification (*e.g.* ATC, CCAM). To query these data, we compared solutions issued from first order logic, Chronicles and MTL. It shows that none of them is enough expressive to correctly *translate* all the desirable constraints of the care pathway. The future work consists of extending these formalisms and to study efficiency issues by testing them through many cases of study on the SNDS.

References

ANICIC D., FODOR P., RUDOLPH S., STÜHMER R., STOJANOVIC N. & STUDER R. (2011). ETALIS: Rule-based reasoning in event processing. In *Proceedings of the Conference on Reasoning in event-based distributed systems*, p. 99–124.

ARTALE A., KONTCHAKOV R., KOVTUNOVA A., RYZHIKOV V., WOLTER F. & ZAKHARYASCHEV M. (2017). Ontology-mediated query answering over temporal data: A survey. In *Proceedings of the International Symposium on Temporal Representation and Reasoning (TIME)*, p. 1–37.

BAADER F., CALVANESE D., MCGUINNESS D., PATEL-SCHNEIDER P. & NARDI D. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge university press.

BAIER C. & KATOEN J.-P. (2008). *Principles of model checking*. MIT press.

BIENVENU M. (2016). Ontology-mediated query answering: harnessing knowledge to get more from data. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 4058–4061.

BOTTRIGHI A., GIORDANO L., MOLINO G., MONTANI S., TERENZIANI P. & TORCHIO M. (2010). Adopting model checking techniques for clinical guidelines verification. *Artificial intelligence in medicine*, **48**(1), 1–19.

BOUYER P., CHEVALIER F. & MARKEY N. (2005). On the expressiveness of tptl and mtl. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, p. 432–443: Springer.

CABALAR P., OTERO R. P. & POSE S. G. (2000). Temporal constraint networks in action. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, p. 543–547.

CLARKE JR E. M., GRUMBERG O., KROENING D., PELED D. & VEITH H. (2018). *Model checking*. Springer.

COMBI C., KERAVNOU-PAPAILIOU E. & SHAHAR Y. (2010). *Temporal information systems in medicine*. Springer.

DELLUC A., IANOTTO J.-C., TROMEUR C., DE MOREUIL C., COUTURAUD F., LACUT K., LE MOIGNE E., LOUIS P., THEREAUX J., METGES J.-P. *et al.* (2018). Real-world incidence of cancer following a first unprovoked venous thrombosis: Results from the EPIGETBO study. *Thrombosis research*, **164**, 79–84.

DOUSSON C. & LE MAIGAT P. (2007). Chronicle recognition improvement using temporal focusing and hierarchization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 324–329.

GIATRAKOS N., ARTIKIS A., DELIGIANNAKIS A. & GAROFALAKIS M. (2017). Complex event recognition in the big data era. *Proceedings of Conference on Very Large Data Base Endowment (VLDB Endow.)*, **10**(12), 1996–1999.

LEVESQUE H. J. (1986). Knowledge representation and reasoning. *Annual review of computer science*, **1**(1), 255–287.

LONG D. (1989). A review of temporal logics. *The Knowledge Engineering Review*, **4**(2), 141–162.

OGER E. & EPI-GETBO (2000). Incidence of venous thromboembolism: a community-based study in western france. *Thrombosis and haemostasis*, **83**(05), 657–660.

OUAKNINE J. & WORRELL J. (2005). On the decidability of metric temporal logic. In *20th Annual IEEE Symposium on Logic in Computer Science (LICS'05)*, p. 188–197.

O'CONNOR M. J., SHANKAR R. D., PARRISH D. B. & DAS A. K. (2009). Knowledge-data integration for temporal reasoning in a clinical trial system. *International journal of medical informatics*, **78**, 77–85.

RIVAULT Y., DAMERON O. & LE MEUR N. (2019). queryMed: Semantic web functions for linking pharmacological and medical knowledge to data. *Bioinformatics*, p. to appear.

SNODGRASS R. T. (1992). Temporal databases. In *Theories and methods of spatio-temporal reasoning in geographic space*, p. 22–64. Springer.

TUPPIN P., RUDANT J., CONSTANTINOU P., GASTALDI-MÉNAGER C., RACHAS A., DE ROQUEFEUIL L., MAURA G., CAILLOL H., TAJAHMADY A., COSTE J. *et al.* (2017). Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'assurance maladie (sniiram) to the système national des données de santé (snds) in france. *Revue d'epidemiologie et de sante publique*, **65**, S149–S167.

WEILL A., PAÏTA M., TUPPIN P., FAGOT J.-P., NEUMANN A., SIMON D., RICORDEAU P., MONTASTRUC J.-L. & ALLEMAND H. (2010). Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. *Pharmacoepidemiology and drug safety*, **19**(12), 1256–1262.

Intégration de connaissances médicales au sein d'un algorithme de classification automatique : application au codage du diabète

Arnaud Serret-Larmande[1], Jean-Baptiste Escudie[1], Catherine Duclos[12]

[1] HÔPITAL AVICENNE, Bobigny, France
arnaud.serret-larmande@aphp.fr
catherine.duclos@aphp.fr
jean-baptiste.escudie@aphp.fr

[2] INSERM UMRS 1142,
Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, Paris, France
catherine.duclos@aphp.fr

Résumé : La plus-value de l'adjonction de connaissances médicales sous formes de règles à un algorithme d'apprentissage automatique est évaluée ici au travers d'un cas d'étude : l'assignation de codes diagnostics de la Classification Internationale des Maladies (CIM-10) à des séjours de diabétologie à partir de documents textuels. La méthodologie hiérarchique développée ici entend simplifier la tâche de prédiction des algorithmes grâce à l'exploitation de la structure hiérarchique de la CIM-10. Les résultats montrent une augmentation des performances de prédiction de près de 10 points de F-score en moyenne pour la méthode hiérarchique comparativement à une méthode traitant chacun des codes de façon indépendante. L'utilisation à plus grande échelle d'une telle approche reste à explorer, et pourrait passer par l'exploitation de terminologies intégrant des relations conceptuelles plus détaillées que pour la CIM-10.

Mots-clés : Apprentissage automatique, expertise médicale, codage hospitalier, CIM-10

1 Introduction

À l'heure où les fruits des recherches en intelligence artificielle en médecine commencent à être utilisé dans la pratique courante (Abràmoff *et al.*, 2018) (Kalyanpur & Murdock, 2015), la question de la place du médecin et de son expertise médicale dans cette révolution en marche est amenée à se poser de façon continue dans un futur proche (Wartman & Combs, 2019). Malgré l'importance de la question, la difficulté pour intégrer cette expertise à la conception de systèmes d'apprentissage automatique rend le sujet délicat à traiter.

Les performances des applications basées sur de l'intelligence artificielle dans le domaine médical varient selon la tâche, et si en imagerie des algorithmes obtiennent d'ores-et-déjà des performances diagnostiques supérieures à celles de radiologues pour quelques cas d'usages (Liu *et al.*, 2018) (Steiner *et al.*, 2018), les applications à d'autres domaines médicaux n'ont pas encore montré d'amélioration substantielle, notamment dans le cas du codage hospitalier (Stanfill *et al.*, 2010 Nov Dec) (Sheikhalishahi *et al.*, 2019) (Topol, 2019). La complexité intrinsèque du domaine médical peut donc apparaître aujourd'hui comme rédhibitoire pour certaines tâches de classification. Ainsi, formaliser des connaissances médicales dans un format exploitable informatiquement pourrait représenter une réponse à ces difficultés.

Pour illustrer ce propos, nous avons souhaité évaluer l'apport de connaissances médicales formalisées à un algorithme de classification en comparant cette approche à un algorithme naïf, sur un cas d'usage précis : l'assignation de codes diagnostics de la Classification Internationale des Maladies (CIM-10) à des comptes rendus médicaux de diabétologie.

L'automatisation du codage des séjours hospitaliers par des algorithmes d'apprentissage est un champ de recherche important, tant du côté universitaire qu'industriel (Xie & Xing, 2018). Les enjeux sont en effet multiples, le codage des séjours hospitaliers n'étant plus seulement utilisé comme base du financement des établissements de soins, mais exploité de plus en plus fréquemment notamment pour l'évaluation de la qualité des soins, ou dans le cadre de la recherche biomédicale, favorisé notamment par les avancées récentes en intelligence

artificielle (Daien *et al.*, 2017) (Rajkomar *et al.*, 2018). Cependant, les performances des diverses approches réalisées pour assigner automatiquement des codes diagnostics aux séjours hospitaliers se sont révélées insuffisantes jusqu'à présent pour envisager leur utilisation en pratique courante.

Une approche naïve consiste à considérer les codes diagnostics issus de la CIM-10 comme un ensemble de labels indépendants, et a entraîner un algorithme évaluant la probabilité de chacun de ces codes d'être assigné à un séjour donné. Or la CIM-10 est organisée de façon hiérarchique en suivant une succession d'embranchements, depuis les chapitres généraux représentant des grands groupes nosologiques jusqu'aux codes diagnostics terminaux. Plusieurs codes peuvent ainsi partager une partie de leur signification respective en fonction de leur proximité dans cette arborescence. Quelques auteurs ont tenté d'exploiter cette structure hiérarchique de façon automatisée, par exemple via l'utilisation de réseaux de neurones convolutionnels dessinés pour apprendre la structure hiérarchique (Catling *et al.*, 2018), ou via une approche pas-à-pas descendante (Perotte *et al.*, 2014). Dans ces deux cas, l'utilisation de l'approche hiérarchique a permis d'améliorer les performances des modèles.

Nous avons souhaité ici évaluer une méthodologie mettant à profit la hiérarchie et les similitudes nosologiques existants parmis les codes diagnostics issus de la CIM-10 afin d'améliorer les performances de classification de sept algorithmes d'apprentissage automatique. Cette méthodologie développée spécifiquement pour cette étude sera dénommée méthode "hiérarchique", par opposition à l'approche naïve traitant l'ensemble des codes diagnostics comme indépendants, ci-après dénommée "indépendante".

2 Matériel et méthodes

2.1 Données et labels

Les données utilisées pour cette tâche de classification sont les comptes rendus médicaux, notes cliniques et ordonnance associés aux séjours du service de diabétologie de l'hôpital Avicenne (Bobigny, France), enregistrés entre le 1er septembre 2017 et le 1er février 2019. Le critère d'inclusion principal était la présence d'au moins un code CIM-10 appartenant aux sous-chapitres diabète de type 1 (E10.) ou 2 (E11.) en tant que diagnostic principal, diagnostic relié ou diagnostic associé secondaire. Cet ensemble de documents a été choisi pour deux principales raisons. Après application de ces critères de sélection, 1049 séjours ont été retenus.

Type de diabète		Type1	Type 2	
Insulino-dépendance		Oui	Oui	Non
Complications	Coma	E100	E1100	E1108
	Acidocétose	E101	E1110	E1118
	Rénale	E102	E1120	E1128
	Oculaire	E103	E1130	E1138
	Neurologique	E104	E1140	E1148
	Vasculaire	E105	E1150	E1158
	Autres précisées	E106	E1160	E1168
	Aucune complication	E109	E1190	E1198

Tableau 1 – Codes diagnostics des sous-chapitres diabète de type 1 et 2

Les labels retenus pour cette tâche de classification sont les codes issus de la CIM-10 adaptée par l'ATIH (Agence Technique de l'Information sur l'Hospitalisation) pour le PMSI français, appartenant à l'une des subdivisions des branches diabète de type 1 ou diabète de type 2 (Tableau 1). Après exclusion des codes interdits pour le codage des pathologies en lien avec le diabète d'après les recommandations de l'ATIH, 24 codes terminaux faisaient finalement partie des labels pouvant être assignés aux séjours hospitaliers.

Les séjours présentant des critères de mauvaise qualité de codage présence de codes mutuellement incompatibles (concommittance de 2 codes de diabète de type 1 et de type 2concernant le type du diabète, concommitance de 2 codes de diabète sans complication et avec complication, concommitance de 2 code de diabète insulino requérant et non insulino requérant), ou codes interdits pour le codage du diabète (omplications multiples et complications non précisées) ont été exclus. Après application de ces critères d'exclusions, 977 séjours ont finalement été retenus pour l'analyse (voir Figure 2).

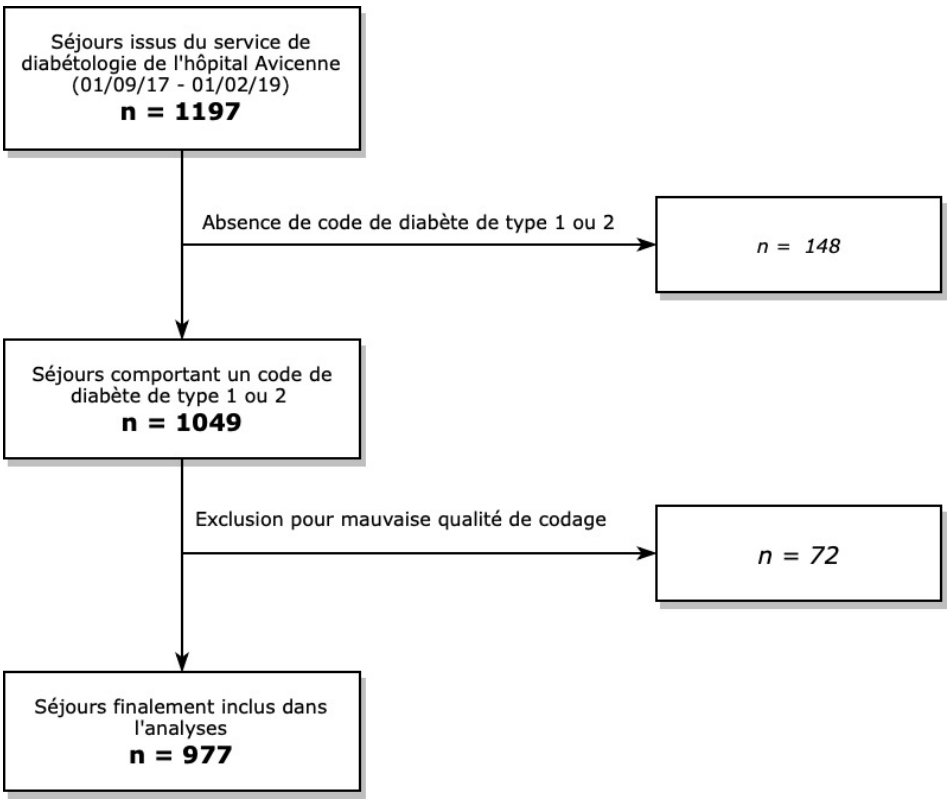


FIGURE *1 – Flow-chart : sélection des séjours*

2.2 Méthodologies de prédiction

L'assignation de codes CIM-10 à des comptes rendus de séjours hospitaliers correspond à une tâche de classification multi-catégorielle et multi-label : un séjour peut se voir attribuer un ou plusieurs codes diagnostics, et ceux-ci ne sont pas mutuellement exclusifs (par exemple l'ensemble de code [E101, E104, E105] peut être codé pour un même séjour) .

Concernant la méthodologie indépendante, le système devait être capable de prédire chacun des codes indépendamment les uns des autres. Nous avons ainsi entraîné un ensemble de classifieurs binaires, un pour chaque code présent dans le gold-standard de notre jeu de données. Les prédictions de ces classifieurs étaient *in fine* regroupées pour donner l'ensemble de codes assignés à un séjour.

Concernant la méthodologie hiérarchique, les labels terminaux ont été décomposés en plusieurs parties en fonction de leur signification, ce qui consistait ici à subdiviser chaque code en trois parties : une première pour le type du diabète, une deuxième pour les différentes complications liées au diabète et finalement une troisième pour la présence ou l'absence d'une requérance à l'insuline (voir Figure 2). Le système était conçu pour prédire spécifiquement chacune de ces sous-parties, à l'aide de multiples classifieurs binaires. Les prédictions étaient ensuite regroupées et les différents codes reconstitués à partir de la combinaison de chacune

de ces prédictions. Par exemple, si le système prédisait un diabète de type 2 (E11.), avec des complications oculaires (.2), rénales (.3) et vasculaires périphériques (.5) ainsi qu'une absence de requérance à l'insuline (..8), la combinaison de ces prédictions donnait l'ensemble de codes suivant : [E1128, E1138, E1158].

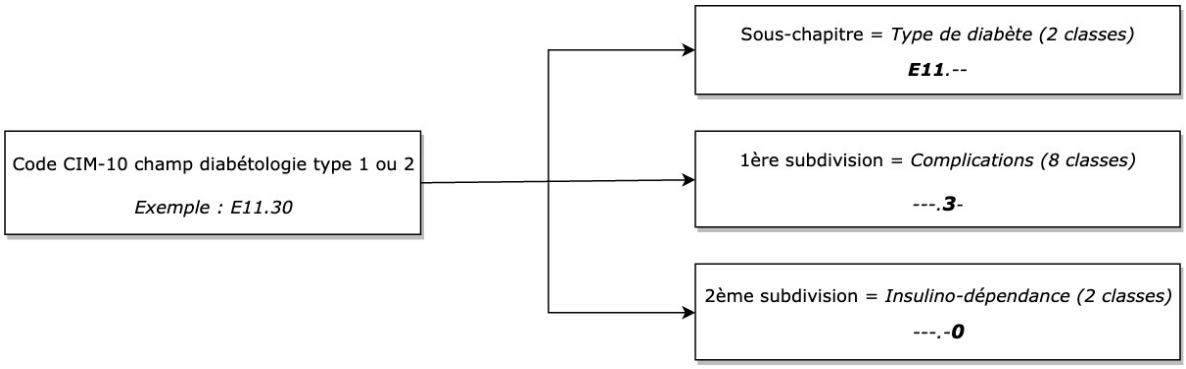


FIGURE *2 – Décomposition des codes terminaux par la méthodologie hiérarchique*

Cette décomposition de la tâche basée sur le sens médical des codes représentait l'expertise médicale apportée à l'algorithme : la décomposition des codes à prédire suivi d'une reconstruction à partir des prédictions permettait de réduire l'espace de prédictions possible et au passage de s'affranchir de la possibilité de prédiction de codes incompatibles. En effet, là où une méthodologie standard traitant chaque code de façon indépendante peut au final prédire des combinaisons de codes aberrantes pour un même séjour, telles que E10.1 (diabète de type 1 avec acidocétose) et E11.90 (diabète de type 2 insulino-dépendant sans complications), la reconstruction de codes à partir de prédictions intermédiaires permet d'éviter cet écueil. Secondairement, décomposer les codes de cette façon permettait d'obtenir plus d'exemples d'entraînement pour chacun des labels à prédire. En effet la décomposition en label intermédiaire permettait de regrouper tous les exemples positifs pour une même complication ou un même type de diabète ensemble, ce qui présentait un intérêt dans le cas de code décrivant des diagnostics dont la prévalence est très faible, et ce indépendemment de la taille du jeu d'entraînement. Par exemple, le code E1118 diabète de type 2 avec acidocétose non dépendant à l'insuline est peu fréquent de par la rareté même d'une telle pathologie, bien que les composants diabètes de type 2, acidocétose et dépendance à l'insuline ne soient pas spécifiquement rare pris isolément.

2.3 Algorithmes de classifications et pré-traitement des données

Les prédictions ont été réalisées à partir de l'ensemble des documents textuels disponibles pour un séjour. Chaque mot a été tokénizé et encodé selon son indice Term-Frequency/Inverse-Document-Frequency.

Le jeu de données a été divisé en 2 parties, avec 80% des données pour le jeu d'entraînement et 20% des données pour le jeu de test qui a servi à l'évaluation des performances.

Pour chacune des deux méthodologies présentées, les algorithmes d'apprentissage suivants ont été testés : Régression Logistique avec L1-pénalisation (Marquardt & Snee, 1975), Decision Tree, Random Forest, AdaBoost - Logistic Regression (Freund & Schapire, 1997), AdaBoost - Decision Tree (Freund & Schapire, 1997), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP). L'optimisation des hyperparamètres pour chacun de ces algorithmes a été réalisée selon l'approche *random search*, celle-ci pouvant converger vers un ensemble d'hyperparamètres optimal plus rapidement que la méthode *grid search* (Bergstra & Bengio, 2012). De plus un algorithme de classification naïf (Dummy Classifier) a été utilisé afin d'avoir une comparaison de base, celui-ci prédisant pour un séjour chaque label en fonction de la probabilité *a priori* de ce label (prévalence du label dans les données d'entraînement), indépendamment des documents textuels associés à ce séjour.

2.4 Evaluation

Les performances en matière de prédictions ont été évaluées selon trois métriques : la précision (ou valeur prédictive positive, rapport vrais positifs par nombre d'exemples prédits comme positifs), le rappel (ou sensibilité, rapport vrais positifs par nombre d'exemples réellement positifs) et le F-score (moyenne harmonique de la précision et du rappel). Ces métriques ont été retenues en se basant sur la littérature de référence (Koyejo *et al.*, 2014). Etant donné la nature multi-catégorielle et multi-label de la tâche de prédiction, un indicateur synthétique a été obtenu pour chacune des trois métriques par agrégation du nombre total de vrais positifs, faux positifs, vrais négatifs et faux négatifs pour chacun des labels – micro-averaging (Perotte *et al.*, 2014).

Les performances de prédiction ont été évaluées pour la prédiction des labels terminaux, à savoir les codes CIM-10 utilisés pour le codage des diagnostics, mais également pour la prédiction des labels intermédiaires construits pour l'implémentation de la méthode hiérarchique (et pouvant être déduits des prédictions de la méthodologie indépendante).

2.5 Logiciel

L'ensemble du code pour ce travail a été développé en python 3.6. Les librairies Scikit-learn 0.20 (Pedregosa *et al.*, 2011) et NLTK 3.4 (Bird *et al.*, 2009) ont été utilisées pour le pré-traitement des données, et Scikit-learn 0.20 pour l'implémentation des algorithmes et la recherche d'hyperparamètres.

3 Résultats

3.1 Description de la population

Au total, les 977 séjours inclus rassemblaient 3761 documents.

La répartition des labels est donnée par le tableau 2. La première ligne donne le nombre d'occurrences du code diagnostic dans l'ensemble du corpus, la deuxième la fréquence rapportée au nombre de séjours. Ainsi, la distribution des codes apparaît très déséquilibrée, avec une fréquence moyenne par code et par séjour de 8,5% (en excluant les codes complications "0", désignant les diabètes compliqués de coma, pris en charge en réanimation et non en service de diabétologie), mais pouvant aller de moins de 1% pour les codes E105 (diabète de type 1 avec complications vasculaires périphériques et E1118 (diabète sucré de type 2 non insulinotraité avec acidocétose), à plus de 35% pour le code E1120 (diabète de type 2 insulinotraité avec complications rénales).

Codes CIM-10	E100	E101	E102	E103	E104	E105	E106	E109	E1100	E1108	E1110	E1118
Nombre d'occurrences	0	39	40	68	52	1	21	60	0	0	27	1
Fréquence	0	0,04	0,04	0,07	0,05	0,	0,02	0,06	0	0	0,03	0
Codes CIM-10	E1120	E1128	E1130	E1138	E1140	E1148	E1150	E1158	E1160	E1168	E1190	E1198
Nombre d'occurences	343	31	315	25	326	37	24	6	179	22	110	30
Fréquence	0,35	0,03	0,32	0,03	0,33	0,04	0,02	0,01	0,18	0,02	0,11	0,03

Tableau 2 – Fréquence des différents codes par séjours (première ligne = nombre d'occurence dans corpus, deuxième ligne fréquence rapportée au nombre de séjours)

3.2 Prédictions des codes terminaux

Les performances de prédiction des labels terminaux sont résumées dans le tableau 3. La performance la plus élevée de prédiction globale est obtenue par la méthodologie hiérarchique avec pour classifieur la régression logistique pénalisée (F-score = 0,576). En utilisant la méthodologie indépendante, la performance la plus élevée est obtenue en utilisant un SVM pour estimateur, mais avec une performance relativement inférieure (F-score = 0,493). Pour

Algorithme	Méthodologie	F-score	Précision	Rappel
Logistic Regression	Hiérarchique	**0,576**	**0,587**	**0,565**
	Indépendante	0,481	0,451	0,515
Random Forest	Hiérarchique	0,574	**0,599**	0,550
	Indépendante	0,472	0,390	**0,597**
AdaBoost - Decision Tree	Hiérarchique	**0,566**	**0,574**	**0,559**
	Indépendante	0,467	0,445	0,491
SVM	Hiérarchique	**0,545**	0,551	**0,538**
	Indépendante	0,493	**0,560**	0,441
AdaBoost - Logistic Regression	Hiérarchique	**0,538**	**0,563**	0,515
	Indépendante	0,387	0,290	**0,582**
MLP	Hiérarchique	**0,511**	0,501	**0,521**
	Indépendante	0,411	**0,664**	0,297
Decision Tree	Hiérarchique	0,177	0,161	0,197
	Indépendante	**0,330**	**0,235**	**0,550**
Dummy Classifier	Hiérarchique	0,211	0,214	0,209
	Indépendante	0,212	0,202	0,224

Tableau 3 – Performance pour la prédiction des codes terminaux

chacun des algorithmes d'apprentissage testés, les performances semblent significativement meilleures avec l'utilisation de la méthodologie de prédiction dite hiérarchique (F-score augmenté de 9 points ou plus pour 5 des 7 algorithmes testés).

S'il apparaît une tendance nette à l'amélioration du F-score grâce à la méthodologie hiérarchique (hormis avec l'utilisation du Decision Tree), l'effet sur la précision et le rappel n'est pas uniforme et l'un ou l'autre de ces indicateurs peut être augmenté ou diminué selon les différents algorithmes.

Concernant la prédiction label par label les performances sont grandement influencées par la prévalence du code dans le jeu de donnée, mais on ne retrouve pas d'amélioration nette de la performance pour les labels peu fréquents en utilisant la méthodologie hiérarchique (résultats non montrés).

3.3 Performances dans la prédiction des codes intermédiaires

Les performances de prédictions pour les labels intermédiaires sont indiquées dans le tableau 4 (seul l'algorithme ayant les meilleurs performances de prédiction les labels terminaux est reporté). Les performances de prédiction apparaissent substantiellement améliorées pour chacun des trois labels intermédiaires, dans des proportions plus importantes pour le type de diabète et les différentes complications que pour l'insulino-requérance.

4 Discussion

Ainsi, la méthodologie hiérarchique implémentée ici a obtenu une augmentation moyenne de près de 10 points de F-score pour les algorithmes les plus performants.

Deux hypothèses théoriques semblaient susceptibles d'améliorer les performances de prédiction de la méthodologie hiérarchique : l'augmentation du nombre d'exemples d'entraînement par classifieur et la simplification de la tâche de prédiction. Les résultats code par code montraient que les codes les moins fréquents (donc les plus susceptibles de bénéficier d'une

Niveau	Algorithme	Méthodologie	F-score	Précision	Rappel
Complications	Logistic Regression	Hiérarchique	**0,696**	**0,709**	**0,682**
		Indépendante	0,618	0,606	0,629
	Dummy Classifier	Hiérarchique	0,360	0,364	0,356
		Indépendante	0,327	0,319	0,335
Insulino-dépendance	Logistic Regression	Hiérarchique	**0,803**	**0,795**	**0,810**
		Indépendante	0,717	0,665	0,778
	Dummy Classifier	Hiérarchique	0,627	0,607	0,647
		Indépendante	0,631	0,554	0,732
Type diabète	Logistic Regression	Hiérarchique	**0,903**	**0,903**	**0,903**
		Indépendante	0,779	0,794	0,765
	Dummy Classifier	Hiérarchique	0,704	0,704	0,704
		Indépendante	0,698	0,668	0,730

Tableau 4 – Performance pour la prédiction des codes intermédiaires

augmentation du nombre d'exemples d'entraînement) ne semblaient pas prédits avec une plus grande précision. Ainsi, l'augmentation globale des performances pourrait être attribuée au moins en partie à la simplification de la tâche de classification.

Un autre aspect bénéfique de cette implémentation hiérarchique pourrait être la plus grande proximité des codes prédits avec les codes réels. En effet, en regardant les performances dans la prédiction des labels intermédiaires, l'amélioration la plus importante est obtenue dans la prédiction du type de diabète (+12 points de F-score pour la régression logistique), c'est-à-dire pour le label intermédiaire représentant le noeud le plus en amont de l'arborescence de la CIM-10. Ainsi, dans le cas où la prédiction terminale s'avérerait fausse, les codes prédits pourraient se trouver tout de même plus proche du gold-standard, élément important dans l'optique d'une utilisation en pratique courante.

L'une des limitations de l'étude est l'utilisation de méthodes qui ne représentent pas l'état de l'art dans le champ du traitement automatique de la langue et de l'intelligence artificielle. En effet, l'encodage utilisé faisait appel à la méthode dite Bag-Of-Words (sac de mots), et non à l'utilisation d'embeddings qui constituent actuellement la méthode de référence pour l'encodage de texte en prenant en compte le contexte de chaque mot (Young *et al.*, 2018). Par ailleurs les algorithmes utilisés n'incluent pas de réseaux de neurones convolutionnels ou récurrents, qui sont les algorithmes de référence utilisés pour des tâches de prédictions à partir de documents textuels (Young *et al.*, 2017). La principale raison à cela était la taille limitée du corpus d'entraînement, celui-ci ayant été spécifiquement choisi en raison de la plus grande qualité des labels. Ce relativement faible nombre d'exemples ne permettait pas d'utiliser des modèles d'apprentissage profonds, dont les performances sur des corpus de tailles réduites sont notoirement limitées en raison du surapprentissage. Cependant, il nous semble que ces limitations ne remettent pas en question la validité des résultats présentés. D'une part, les performances d'algorithmes standards peuvent soutenir la comparaison dans certaines tâches, même comparés à des réseaux de neurones profonds (Rajkomar *et al.*, 2018), et d'autre part, en temps qu'étude de cas, nous souhaitions étudier l'impact de l'introduction de connaissances médicales pour améliorer les performances d'algorithmes d'apprentissage, indépendamment de leurs performances propres. Les résultats soutiennent ce constat : en effet, la variance en termes de qualité de prédiction dans l'éventail des différents algorithmes testés est moindre que la différence, pour chacun des algorithmes, entre les performances de la méthodologie hiérarchique et de la méthodologie indépendante (à l'exception de l'algorithme Decision Tree).

Une autre limitation de l'étude semble être la difficulté pour généraliser ce type de méthode à une tâche de prédiction plus large, idéalement incluant l'ensemble des codes diagnostics

possibles. S'il existe de nombreux chapitres et classes de la CIM-10 qui apparaissent éligibles à l'implémentation d'une méthodologie de prédiction hiérarchique, tels que la classe I60-I69 concernant les maladies cérébrovasculaires (combinaison d'un mécanisme physiopathologique et d'une localisation anatomique), ou encore la classe M15-M19 classant les arthroses (combinant localisations anatomiques et quantification de ces atteintes), il est vrai qu'une potentielle généralisation de cette approche nécessiterait un travail fastidieux pour implémenter une méthodologie hiérarchique pour un plus grand nombre de codes. D'autres approches ont tenté de mettre à profit la hiérarchie existante au sein de la CIM-10 en évitant d'avoir à implémenter un ensemble de règles manuelles (Catling *et al.*, 2018) (Perotte *et al.*, 2014), sans pour autant obtenir des performances suffisantes permettant d'envisager leur utilisation en pratique. Une solution à ce problème pourrait être l'utilisation de systèmes d'organisation des connaissances en santé plus spécialisés qui intègrent une composante relationnelle entre les pathologies de façon plus poussée que la CIM-10, permettant de concilier prise en compte des relations conceptuelles entres les pathologies et utilisation à grande échelle. On peut citer la terminologie SNOMED-CT qui implémente des relations entre phénomènes physiopathologiques et entités anatomiques, ou la 11ème version de la classification CIM (CIM-11) qui intègre la notion de parents multiples pour un même code diagnostic.

En conclusion, ce travail entend mettre en lumière les améliorations en terme de performances offertes lorsqu'un raisonnement médical est intégré dans la conception de systèmes de prédiction basés sur des algorithmes d'intelligence artificielle. L'utilisation d'une telle approche pour un ensemble de codes diagnostics exhaustif pourrait nécessiter de passer par des terminologies plus spécialisées intégrant des relations conceptuelles dans leur conception.

Références

ABRÀMOFF M. D., LAVIN P. T., BIRCH M., SHAH N. & FOLK J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, **1**(1), 39.

BERGSTRA J. & BENGIO Y. (2012). Random Search for Hyper-Parameter Optimization. p.̃25.

BIRD S., LOPER E. & KLEIN E. (2009). *Natural Language Processing With Python*. O'reilly media inc. edition.

CATLING F., SPITHOURAKIS G. P. & RIEDEL S. (2018). Towards automated clinical coding. *International Journal of Medical Informatics*, **120**, 50–61.

DAIEN V., KOROBELNIK J.-F., DELCOURT C., COUGNARD-GREGOIRE A., DELYFER M. N., BRON A. M., CARRIÈRE I., VILLAIN M., DAURES J. P., LACOMBE S., MARIET A. S., QUANTIN C. & CREUZOT-GARCHER C. (2017). French Medical-Administrative Database for Epidemiology and Safety in Ophthalmology (EPISAFE) : The EPISAFE Collaboration Program in Cataract Surgery. *Ophthalmic Research*, **58**(2), 67–73.

FREUND Y. & SCHAPIRE R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139.

KALYANPUR A. & MURDOCK J. W. (2015). Unsupervised Entity-Relation Analysis in IBM Watson. (2015), 12.

KOYEJO O. O., NATARAJAN N., RAVIKUMAR P. K. & DHILLON I. S. (2014). Consistent Binary Classification with Generalized Performance Metrics. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. D. LAWRENCE & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 27*, p. 2744–2752. Curran Associates, Inc.

LIU Y., KOHLBERGER T., NOROUZI M., DAHL G. E., SMITH J. L., MOHTASHAMIAN A., OLSON N., PENG L. H., HIPP J. D. & STUMPE M. C. (2018). Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection. *Archives of Pathology & Laboratory Medicine*.

MARQUARDT D. W. & SNEE R. D. (1975). Ridge Regression in Practice. *The American Statistician*, **29**(1), 3–20.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY É. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PEROTTE A., PIVOVAROV R., NATARAJAN K., WEISKOPF N., WOOD F. & ELHADAD N. (2014). Diagnosis code assignment : Models and evaluation metrics. *Journal of the American Medical*

Informatics Association : JAMIA, **21**(2), 231–237.

RAJKOMAR A., OREN E., CHEN K., DAI A. M., HAJAJ N., HARDT M., LIU P. J., LIU X., MARCUS J., SUN M., SUNDBERG P., YEE H., ZHANG K., ZHANG Y., FLORES G., DUGGAN G. E., IRVINE J., LE Q., LITSCH K., MOSSIN A., TANSUWAN J., WANG D., WEXLER J., WILSON J., LUDWIG D., VOLCHENBOUM S. L., CHOU K., PEARSON M., MADABUSHI S., SHAH N. H., BUTTE A. J., HOWELL M. D., CUI C., CORRADO G. S. & DEAN J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, **1**(1), 18.

SHEIKHALISHAHI S., MIOTTO R., DUDLEY J. T., LAVELLI A., RINALDI F. & OSMANI V. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases : Systematic Review. *JMIR medical informatics*, **7**(2), e12239.

STANFILL M. H., WILLIAMS M., FENTON S. H., JENDERS R. A. & HERSH W. R. (2010 Nov-Dec). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association : JAMIA*, **17**(6), 646–651.

STEINER D., MACDONALD R., LIU Y., TRUSZKOWSKI P., HIPP J., GAMMAGE C., THNG F., PENG L. & STUMPE M. (2018). Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *The American Journal of Surgical Pathology*, **42**(12), 1636–1646.

TOPOL E. J. (2019). High-performance medicine : The convergence of human and artificial intelligence. *Nature Medicine*, **25**(1), 44.

WARTMAN S. A. & COMBS C. D. (2019). Reimagining Medical Education in the Age of AI. *AMA Journal of Ethics*, **21**(2), 146–152.

XIE P. & XING E. (2018). A Neural Architecture for Automated ICD Coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1066–1076, Melbourne, Australia : Association for Computational Linguistics.

YOUNG T., HAZARIKA D., PORIA S. & CAMBRIA E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. *arXiv :1708.02709 [cs]*.

YOUNG T., HAZARIKA D., PORIA S. & CAMBRIA E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, **13**(3), 55–75.

Apports de lIntelligence Artificielle à la prédiction des durées de séjours hospitaliers

Rachda Naila Mekhaldi[1], Patrice Caulier[1], Sondes Chaabane[1], Sylvain Piechowiak[1], Abdelahad Chraibi[2]

[1] LAMIH UMR CNRS 8201, Laboratoire d'Automatique, de Mécanique et d'Informatique Industrielles et Humaines, Valenciennes, France
{rachdanaila.mekhaldi,patrice.caulier,sondes.chaabane, sylvain.piechowiak}@uphf.fr

[2] ALICANTE, SECLIN, FRANCE
abdelahad.chraibi@alicante.fr

Résumé :
Au cours des dernières années, les services hospitaliers ont tenté doptimiser leurs ressources afin daméliorer le rendement du fonctionnement de lhôpital. La durée de séjour (DDS) est lun des indicateurs les plus importants pour évaluer ce rendement. Dans un premier temps, cet article identifie les facteurs qui affectent les DDS dans différents services hospitaliers. Un modèle de ces facteurs est alors proposé. Dans un second temps et afin d'analyser les facteurs du modèle et de prédire les DDS, larticle met en évidence un ensemble de techniques et dapplications d'intelligence artificielle dont la fouille de données et les algorithmes dapprentissage automatique. Les Réseaux de Neurones Artificiels, les Arbres de Décision et les Vecteurs à Support Machine sont souvent utilisés dans ce domaine. L'estimation des durées de séjours hospitaliers aide à augmenter la qualité des soins, doptimiser la gestion des ressources des hôpitaux et, enfin, de diminuer les coûts. Une voie de solution est tracée pour la prédiction des durées de séjours hospitaliers. Un modèle représentant les DDS est donné et une démarche méthodologique est présentée.

Mots-clés : Durées de séjours hospitaliers, exploration de données, fouille de données, apprentissage automatique, modèle de prédiction.

1 Introduction

Les établissements de soins cherchent sans cesse à optimiser le fonctionnement de leurs services tout en assurant la qualité de ces services. La planification des activités et la gestion des ressources ont un impact important sur cet objectif. Lestimation de la durée de séjour dun patient au moment de son admission et durant son séjour est un indicateur dévaluation de base des services de santé. La durée de séjour est définie comme lintervalle de temps entre ladmission du patient et sa sortie (Khosravizadeh *et al.*, 2016). La prédiction des durées de séjours (DDS) hospitaliers contribue à lamélioration de la qualité des soins ainsi qu'à lefficacité de la charge de travail opérationnelle. Cela permet une planification précise des réadmissions, une minimisation des coûts et une réduction du nombre de lits mal occupés. Afin de planifier les activités de soins de manière pertinente, des données de santé sous format électronique sont utilisées pour lanalyse de la variable DDS.

La DDS constitue un indicateur très important dévaluation des performances des hôpitaux, plusieurs facteurs ont une incidence significative sur la DDS (Carter & Potts, 2014). Lobjectif de ce papier est, dune part, de recenser les différents modèles de DDS existants dans un environnement hospitalier puis d'en déduire un modèle généralisé. Dautre part ce papier vise à, mettre en place les méthodes dintelligence artificielle dont la fouille de données et lapprentissage automatique pour lextraction des connaissances à partir des données médicales ainsi que la prédiction des DDS hospitaliers. L'objectif étant de revoir la littérature dans ce contexte afin de cerner les méthodes utilisées dans les deux phases d'extraction de connaissances et d'apprentissage automatique. Pour la phase de prédiction, les algorithmes d'apprentissage supervisé sont employés. En effet, la collecte de ces données auprès des établissements de santé contribue à améliorer les connaissances pour une meilleure prise en charge des patients et a élargi les champs de recherche dans la santé. Néanmoins, il existe

des règles de confidentialité à respecter pour accéder à ces données. Selon le site officiel de l'Agence de l'Information sur l'Hospitalisation (ATIH) un accord doit être établi pour tout titulaire de la part de la commission nationale de linformatique et des libertés (Cnil) (ATIH, 2013). Le dossier d'un patient existe sous format électronique. Il est nommé DSE (Dossier de Santé Electronique). Chaque DSE est créé, géré et conservé par un organisme de santé et seuls les professionnels de santé qui participent aux soins dun patient peuvent y accéder (Tom Seymour, Dean Frantsvog, 2012). Les données du DES concernent des informations générales (identifiant du patient, son âge, genre, situation familiale), des diagnostics, les traitements du patient et son historique médical. Dautres formes de données peuvent porter sur des informations administratives, des modalités dassurance, *etc*. Ces éléments sont hétérogènes et proviennent de multiples sources. Ce qui rend leur traitement complexe. De plus, les bases de données de santé qui concernent les séjours hospitaliers sont souvent volumineuses et contiennent plusieurs données manquantes. Ces dernières ne sont pas disponibles au moment de l'admission du patient.

L'objectif de l'étude est de présenter, différents modèles de durées de séjours hospitaliers existants dans la littérature puis de déduire un modèle généralisé. Ensuite, nous exposons les méthodes et techniques employées pour la prédiction de cette DDS et nous proposons une démarche méthodologique générique pour la prédiction des DDS dans un environnement hospitalier.

2 Facteurs influençant les DDS

L'estimation des DDS utilisera des informations liées à l'hospitalisation des patients pour chaque séjour effectué. Des données de type PMSI ou Programme de Médicalisation des Systèmes d'Information qui reflète la description et la mesure médico-économique de l'activité hospitalière (E.Faure, 2019). Ce programme permettra d'organiser les allocations budgétaires pour les établissements de soins. Pouvoir en extraire un modèle de prédiction des durées de séjour permettra de faire face aux nouvelles situations. Afin de construire ce modèle, il est indispensable didentifier tous les éléments qui le constituent. La modélisation des durées de séjour en milieu hospitalier a pris différentes formes. En effet, des recherches antérieures ont tenté de grouper les patients en fonction de leur état de santé, en supposant que chaque maladie est associée à une durée de séjour recommandée (Shea *et al.*, 1995) par conséquent, la durée de séjour est définie différemment d'un service hospitalier à un autre. Dans (Pendharkar & Khurana, 2014), le type du service de lhôpital est considéré comme un paramètre important pour la prédiction de la durée de séjour parmi les facteurs suivants : le nom de lhôpital, le type du service, le nombre de lits libres, la politique de sortie. Nous avons étudié les différents modèles de durées de séjour hospitalier selon un type de service bien défini. Nous avons constaté que les auteurs se focalisent souvent sur les services hospitaliers suivants : le service de chirurgie générale, le service d'urgence, le service de cardiologie et le service de soins intensifs. Nous avons donc restreint notre étude sur ces 3 services dans ce papier. En se qui concerne les autres services, la définition des DDS et leur représentation est éloignée. Pour le service d'urgence par exemple, la durée de séjour est représenté par nombre d'heure et non pas par nombre de jours.

Dans ce qui suit, nous présentons les facteurs qui influencent les DDS par type de service. Divers modèles de DDS sont donc mis en uvre. Un modèle générique de DDS est enfin donné.

2.1 Facteurs influençant les DDS dans un service de cardiologie

Plusieurs travaux sintéressent à la prédiction des DDS dans les services de cardiologie. Dans létude menée par (Lafaro *et al.*, 2015), 8 variables ont été sélectionnées parmi 36 pour concevoir le modèle de prédiction des DDS. Ces variables liées à lunité de soins intensifs chirurgicaux cardiaques sont : lutilisation d'une pompe de ballon intra-aortique, niveau démission de lO2, utilisation dagents isotropes cardiaques positifs, lhématocrite, le taux de sérum créatinine et la mesure danalyse du gaz du sang. Les facteurs utilisés dans (Hachesu *et al.*, 2013) sont la consommation des médicaments anticoagulants et de la nitrate, mesure

de la pression artérielle diastolique et de cholestérol, douleur thoracique, fraction déjection, comorbidité associée, densité de lipoprotéine, si le patient est fumeur ou non et le taux d'hémoglobine. Dans l'étude de (Tsai *et al.*, 2016), ladresse du patient, son diagnostic, le type de lintervention, la comorbidité et le mode de remboursement ont permis de construire le modèle de la prédiction des DDS. Dans (Almashrafi *et al.*, 2016) les auteurs ont analysé les éléments qui influencent une longue durée de séjour dans le service de soins intensifs cardiaque. Ils ont cité lhistorique des maladies du patient, le nombre de complications, la fraction déjection ventriculaire gauche, *etc*. Ces éléments sont disponibles au moment de ladmission du patient. Pour lensemble de ces études, lâge et le sexe du patient sont toujours utilisés.

À partir des éléments décrits ci-dessus, nous avons constaté que la modélisation des DDS dans un service de cardiologie requiert des informations qui ne sont pas disponibles au moment de ladmission du patient. Comme nous avons également constaté quil existe plusieurs facteurs communs dans les différentes études. Essentiellement, ces facteurs sont : les facteurs démographiques, la mesure du taux de créatinine, le taux de l'hématocrite, *etc*. Ces facteurs sont représentés avec des termes différents dans la littérature mais ils portent la même signification. Nous retrouvons par exemple la temperature du corps indiqué comme une mesure prise séparément et celle comprise dans les signes vitaux. De ce fait, l'aide des experts dans le domaine médical est primordial pour avoir un ensemble d'attributs cohérents pour caractériser les DDS dans la phase de sélection de variables.

2.2 Facteurs influençant les DDS dans un service de soins intensifs (SSI)

Plusieurs travaux se sont intéressés à lanalyse des facteurs qui influencent les DDS dans les unités de soins intensifs en général. La base de données MIMIC III est utilisée pour classer les DDS en deux groupes : longue et courte après avoir quitté le SSI. Les facteurs utilisés concernent ladmission, le transfert, la sortie du patient, les prescriptions médicales, les tests du laboratoire, le diagnostic et les facteurs démographiques (Gentimis *et al.*, 2017). Les données concernant 311 patients ont été utilisées dans (Maharlou *et al.*, 2018) pour identifier les facteurs qui impactent les DDS. Selon cette étude, les facteurs démographiques, l'historique médical (maladie cardiaque, rénale, maladie pulmonaire), taux de créatinine, rythme cardiaque, *etc*. constituent ces attributs. Nous avons remarqué que le modèle de DDS dans un SSI est complexe et nécessite une bonne connaissance du domaine de la santé. Étant donné que ce service accueille des patients dans un état de santé difficile, il existe de forte chance d'avoir des complications ce qui va augmenter la DDS.

2.3 Facteurs influençant les DDS dans un service de chirurgie

Différentes études ont ciblé le service de chirurgie pour analyser les variables utilisées caractéristiques des DDS. La prédiction des DDS est utile en phase de préopération ou postopération. Dans (Chuang *et al.*, 2016), les auteurs ont montré que dans un même service de chirurgie, les facteurs sont différents pour une opération urgente et pour une opération non urgente. Ces facteurs sont : les informations démographiques, lhistorique médical du patient, les mesures des signes vitaux et des résultats du laboratoire, et les données sur lopération reportées par le médecin et les infirmiers. Dans (Khosravizadeh *et al.*, 2016), les facteurs qui influencent le plus la DDS sont : la situation familiale du patient, ses conditions de sortie, le type du traitement ainsi que le type de paiement pour les dépenses de lhôpital. Les recherches dans (Aghajani & Kargari, 2016) se sont focalisées sur la sélection des variables suivantes : le type de lopération, le nombre dopérations quun patient a subi, le temps entre lordre de sortie et la sortie effective du patient, les informations de transfert entre services, le nombre moyen de visites par jour, le nombre de consultations médicales, les hospitalisations précédentes du patient et le nombre de tests effectués pour le service de chirurgie générale.

3 Analyse et discussion à propos des facteurs influençant les DDS

Suite à ces recherches, nous avons analysé ces différents facteurs et nous avons conclu ce qui suit :

- La prédiction des durées de séjour hospitalier est déterminée par le type de service étudié.
- Les services de cardiologie sont souvent les plus étudiés. Cela peut être dû au fait que ce service demande une plus grande part de financement de l'hôpital.
- Il existe des facteurs en commun avec les services de cardiologie, de soins intensifs et de chirurgie. Nous retrouvons lhistorique médical du patient, son rythme cardiaque, analyses du laboratoire, *etc*. Les facteurs démographiques tels que lâge, le genre et la situation familiale du patient sont souvent utilisés.
- Pour un modèle plus généralisé et qui ne se concentre pas sur un seul service hospitalier, il est essentiel de se référer aux spécialistes dans le domaine médical afin de nous guider dans notre sélection de variables comme une première étape de notre processus d'étude. Nous avons résumé les facteurs communs qui figurent dans les différents services dans la figure 1.

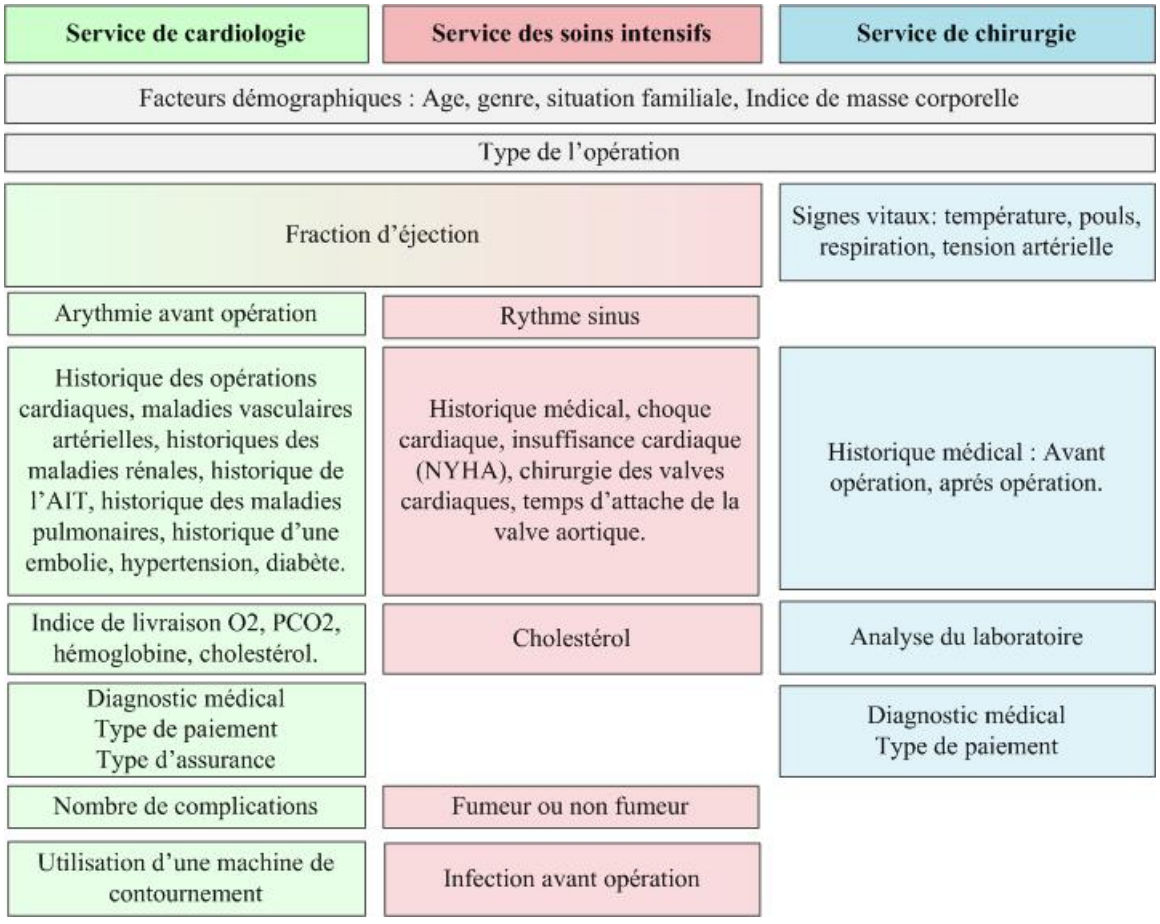


FIGURE *1 – Facteurs impactant les DDS.*

Afin de sélectionner les facteurs qui influencent les DDS et de concevoir un modèle de prédiction des durées de séjour hospitalier, des techniques d'intelligence artificielle ont été utilisées. Dans la section suivante, nous présentons ces techniques et proposons par la suite notre solution.

4 Techniques d'Intelligence Artificielle pour la prédiction des DDS

Lintelligence artificielle a fait un retour fracassant dans le domaine de la santé. Plusieurs algorithmes daide à la décision sont utilisés pour le diagnostic médical, lanalyse des images de radiologie, létude des parcours de patients, *etc* (Amato *et al.*, 2013 ; Pesapane *et al.*, 2018). Le choix des algorithmes d'apprentissage automatique dépend essentiellement de l'objectif de l'étude, de la taille, la qualité et la nature de la base de données. Dans (Liao *et al.*, 2016), les auteurs ont appliqué le clustering sur un ensemble de patients atteints dune maladie rénale en phase terminale initiés à lhémodialyse (HD). Ce groupement est basé sur un modèle de changement des coûts des soins de santé avant et après l'HD selon le profil des patients. Ce modèle utilise deux algorithmes K-means et CAH. Dans une autre étude, lalgorithme K-means est employé pour analyser la variation dans les volumes dentrées et de sorties des hôpitaux et les regrouper selon laffectation des ressources et lefficacité des services de soins (Tseng *et al.*, 2015). Un intérêt important est donné aux méthodes de la fouille de données pour lanalyse des données médicales et dapprentissage automatique pour la classification de ces données. Les données médicales ont été largement utilisées ces dernières années dans le développement des systèmes de soins. Dans le volet de la prédiction des durées de séjour hospitalier, plusieurs travaux existent. Dans (Hachesu *et al.*, 2013), les auteurs ont comparé différents modèles de réseaux de neurones artificiels pour prédire la DDS dans lunité de soins intensifs cardiaque. Une comparaison est établie dans (Shea *et al.*, 1995) entre les réseaux de neurones et les réseaux de neurones à base de système dinférence flous. En plus des réseaux de neurones, les forêts darbres décisionnelles sont implémentées dans (Pendharkar & Khurana, 2014). Les réseaux de neurones, les arbres de décision et les machines à vecteurs de support (SVM) sont utilisés pour la prédiction des durées de séjours des patients souffrants dune coronaropathie (Khosravizadeh *et al.*, 2016). Un modèle de prédiction des longues durées de séjour, avant une opération, basé des arbres de décision, SVM et le random forest a abouti à un meilleur résultat avec le random forest (Almashrafi *et al.*, 2016). Dautres méthodes dapprentissage supervisé sont citées dans (Carter & Potts, 2014) telles que les arbres de régression et de classification (CART) et Chi-Squared Automatic Interaction Detector (CHAID). Les techniques de régression linéaire, lalgorithme naif de Bayes et les k plus proches voisins sont appliquées pour déterminer les facteurs qui influencent les DDS dans un service de chirurgie générale (Gentimis *et al.*, 2017). Plusieurs types darbres de décision et de régression sont exploités dans létude (Tom Seymour, Dean Frantsvog, 2012) afin de décider si une DDS pour un patient après opération est dans un intervalle standardisé et dans (Tsai *et al.*, 2016) pour détecter des situations critiques dans le service durgence de pédiatrie. Les réseaux de neurones sont aussi utilisés dans (Lafaro *et al.*, 2015) pour la prédiction des durées de séjour pour les patients atteints de lune de ces trois maladies : maladie de lartère coronaire, arrêt cardiaque et infarctus aigu du myocarde.

À partir de ces différentes recherches, nous constatons que les méthodes dapprentissage supervisé sont employées pour le problème de prédiction. Il s'agit d'attribuer un séjour hospitalier à une catégorie définie. Les comparaisons entre les différentes méthodes citées auparavant dépendent de l'étude. Généralement les algorithmes random forest et réseaux de neurones ont les meilleures précisions. Malgré les des résultats satisfaisants obtenus à l'aide de ces méthodes, leur faiblesse réside dans le fait que ces méthodes sont utilisées pendant le séjour du patient et non pas juste au moment de son admission.

Pour pallier cette faiblesse, nous proposons une approche de prédiction dynamique de la durée de séjour selon des intervalles de temps séparés. Pour un service donné, une première étape est de définir une durée minimale de séjour représentant le nombre de jours que le patient doit passer dans ce service. Une étape de sélection de variables est réalisée. Des méthodes de fouille de données et de clustering seront explorées pour restreindre l'ensemble de variables représentant les DDS. Ces principales méthodes sont l'algorithme de classification non hiérarchique K-means, les méthodes d'analyse factorielle telle que l'Analyse des Composantes Principales (ACP), les tests statistiques, *etc*. L'intervention des experts dans le domaine médical est nécessaire pour valider les résultats. Pour la phase de classification automatique, un premier modèle de prédiction est réalisé pendant le séjour du patient. Ce modèle concerne un séjour avec une durée minimale fixée auparavant. Ensuite, ce modèle sera enrichi

au fur et à mesure pendant le séjour du patient en collectant plus de données durant le séjour du patient. De ce fait, le modèle proposé étudie des DDS avec une durée minimale d'un jour à l'hôpital. Le choix des intervalles de temps et le type que prendra les DDS dépendront de l'ensemble de données et des expérimentations. Le diagramme de la figure 2 illustre cette approche.

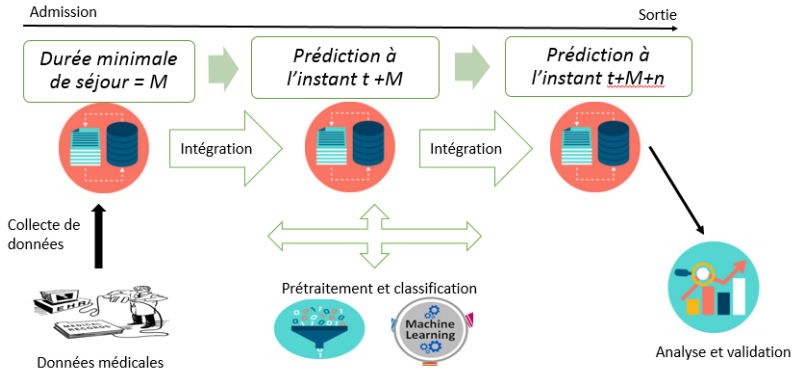


FIGURE 2 – *Prédiction des DDS : proposition.*

Pour mettre en évidence notre solution proposée, des expérimentations vont être menées sur les données médicales du Centre Hospitalier de Valenciennes (CHV). Les données de type PMSI et celles stockées dans les systèmes d'information du CHU seront exploitées.

5 Conclusion

La prédiction des durées de séjour prend de lampleur ces dernières années afin de planifier les activités de soins de manière pertinente. C'est un indicateur de performance de la qualité et de l'efficacité des services des hôpitaux. Dans ce travail, nous avons présenté lutilité de la prédiction des durées de séjour hospitalier en donnant plusieurs modèles de durées de séjour. Cette problématique est complexe car elle dépend de nombreux facteurs. Nous avons mis en évidence les facteurs qui influencent la variable DDS en se basant sur les principaux travaux de recherche effectués dans ce domaine. Ainsi, des solutions se basant sur les méthodes dintelligence artificielle dont la fouille de données et lapprentissage automatique sont exposées. Enfin, une démarche méthodologique générique est proposée pour la prédiction des durées de séjour dans un environnement hospitalier.

Ce papier a ouvert des axes de recherche future. Ces principaux axes sont en premier l'estimation de la durée de séjour au moment de l'admission du patient en vue des données disponibles à cet instant. Deuxièmement, la limite des travaux présentés réside dans la restriction de l'étude à un seul service. Il serai donc important d'élargir le paramètre de l'étude sur plusieurs services voir sur des services ayant des facteurs impactant les DDS communs.

Références

AGHAJANI S. & KARGARI M. (2016). Determining Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department. *Hospital practice and research (HPR)*, **1**(2), 53–58.

ALMASHRAFI A., ALSABTI H., MUKADDIROV M., BALAN B. & AYLIN P. (2016). Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in Oman : A retrospective observational study. *BMJ Open*, **6**(6).

AMATO F., LÓPEZ A., PEÑA-MÉNDEZ E. M., VAHARA P., HAMPL A. & HAVEL J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, **11**(2), 47 – 58.

ATIH (2013). Accès aux données. https ://www.atih.sante.fr/bases-de-donnees/commande-de-bases.

CARTER E. M. & POTTS H. W. (2014). Predicting length of stay from an electronic patient record system : A primary total knee replacement example. *BMC Medical Informatics and Decision Making*, **14**(1), 1–13.

CHUANG M. T., HU Y. H., TSAI C. F., LO C. L. & LIN W. C. (2016). The Identification of Prolonged Length of Stay for Surgery Patients. In *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, p. 3000–3003.

E.FAURE (2019). Le programme de médicalisation des systèmes d'information (PMSI). https ://www.caducee.net/DossierSpecialises/systeme-information-sante/pmsi.asp.

GENTIMIS T., ALNASER A. J., DURANTE A., COOK K. & STEELE R. (2017). Predicting Hospital Length of Stay using Neural Networks on MIMIC III Data. In *IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, p. 1194–1201.

HACHESU P. R., AHMADI M., ALIZADEH S. & SADOUGHI F. (2013). Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. *Healthcare Informatics Research*, **19**(2), 121–129.

KHOSRAVIZADEH O., VATANKHAH S., BASTANI P., KALHOR R., ALIREZAEI S. & DOOSTY F. (2016). Factors affecting length of stay in teaching hospitals of a middle-income country. *Electronic physician*, **8**(10), 3042–3047.

LAFARO R. J., POTHULA S., KUBAL K. P., INCHIOSA M. A., POTHULA V. M., YUAN S. C., MAERZ D. A., MONTES L., OLESZKIEWICZ S. M., YUSUPOV A., PERLINE R. & INCHIOSA M. A. (2015). Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre- Incision Variables. *plos one*, p. 1–19.

LIAO M., LI Y., KIANIFARD F., OBI E. & ARCONA S. (2016). Cluster analysis and its application to healthcare claims data : A study of end-stage renal disease patients who initiated hemodialysis Epidemiology and Health Outcomes. *BMC Nephrology*, **17**(1), 1–14.

MAHARLOU H., NIAKAN KALHORI S. R., SHAHBAZI S. & RAVANGARD R. (2018). Predicting length of stay in intensive care units after cardiac surgery : Comparison of artificial neural networks and adaptive neuro-fuzzy system. *Healthcare Informatics Research*, **24**(2), 109–117.

PENDHARKAR P. C. & KHURANA H. (2014). Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals. *International Journal of Computer Science and Applications*.

PESAPANE F., CODARI M. & SARDANELLI F. (2018). Artificial intelligence in medical imaging : threat or opportunity ? radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, **2**(1), 35.

SHEA S., SIDELI R. V., DUMOUCHEL W., PULVER G., ARONS R. R. & CLAYTON P. D. (1995). Computer-generated informational messages directed to physicians : Effect on length of hospital stay. *Journal of the American Medical Informatics Association*, **2**(1), 58–64.

TOM SEYMOUR, DEAN FRANTSVOG T. G. (2012). Electronic health records (EHR). *American Journal of health Sciences*, **3**(3).

TSAI P. F. J., CHEN P. C., CHEN Y. Y., SONG H. Y., LIN H. M., LIN F. M. & HUANG Q. P. (2016). Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. *Journal of Healthcare Engineering*, **2016**, 11 pages.

TSENG S. F., LEE T. S. & DENG C. Y. (2015). Cluster analysis of medical service resources at district hospitals in Taiwan, 20072011. *Journal of the Chinese Medical Association*, **78**(12), 732–745.