Tentatives de suicide, prédire la récidive avec des techniques d'apprentissage statistique

Philippe Combes[1], Stéphanie Combes[2], Martin Monziols[2]

[1] CH MONTPERRIN, Paris, France
philippe.combes@ch-montperrin.fr

[2] INSEE, Paris, France
stephanie.combes@insee.fr
martin.monziols@ensae-paristech.fr

Résumé : Selon l'institut de veille sanitaire, le suicide est responsable de 10500 décès par an contre 176000 à 200000 hospitalisations pour tentative de suicide. En outre, celles-ci donnent parfois lieu à une ou plusieurs récidives. L'institut estime à 12,8 % le taux de tentatives suivies d'une nouvelle hospitalisation. Le suicide est donc devenu une priorité de santé publique. Dans cette étude et en nous appuyant sur les données collectées lors d'une enquête menée dans le cadre du Plan Régional de Santé Suicide, nous nous concentrerons plus particulièrement sur le risque de récidive, que nous tenterons d'évaluer en utilisant des techniques d'apprentissage statistique (ou *machine learning*). En effet, correctement évalué, celui-ci pourrait apparaître comme une information susceptible de permettre aux centres hospitaliers d'améliorer la prise en charge et le traitement individualisé. Les résultats préliminaires obtenus jusqu'ici montrent qu'il est difficile d'identifier correctement les récidivistes sans surestimer le risque pour les autres patients.

Mots-clés : Récidive, Machine learning, analyse de survie, Tentative de suicide

1 Introduction

Selon l'institut de veille sanitaire, le suicide est responsable de 10500 décès par an contre 176000 à 200000 hospitalisations pour tentative de suicide. En outre, celles-ci donnent parfois lieu à une ou plusieurs récidives. L'institut estime à 12,8 % le taux de tentatives suivies d'une nouvelle hospitalisation. Le suicide est devenu une priorité de santé publique comme en atteste la création de l'Observatoire National du Suicide en 2013. La question est large, c'est une question de santé publique et de politique au long cours mais c'est également une question de prise en charge au niveau individuel. Dans cette étude et en nous appuyant sur les données collectées lors d'une enquête menée dans le cadre du Plan Régional de Santé Suicide, nous nous concentrerons plus particulièrement sur le risque de récidive. Correctement évalué, ce dernier pourrait constituer une information capitale pour l'amélioration de la prise en charge et du traitement proposé par les centres hospitaliers aux patients concernés par ce risque.

Les données de santé correspondent souvent à des relevés d'informations au moment de certains évènements et sont marquées par l'absence d'information entre deux évènements. Les techniques statistiques classiques pour l'analyse de la survie visent donc à étudier les durées entre deux évènements, un point de départ et un évènement ou un évènement et un point final. Ces outils ont été largement appliqués dans d'autres domaines tels que les assurances ou la prédiction des défaillances en ingénierie. Les techniques issues de l'apprentissage statistique mises en œuvre dans cette étude ne sont pas spécifiques à l'analyse de la survie. Utilisées très largement pour la prédiction de variables catégorielles (classification) ou continues (régression), elles sont fondées sur l'automatisation de la sélection de modèles via l'optimisation

d'un critère de qualité/précision de la prédiction. Elles ont, en outre, l'avantage de pouvoir prendre en compte un grand nombre de variables et de paramètres, permettant d'améliorer les performances d'un modèle prédictif par le biais de fonctions de liaison complexes et parfois hautement non linéaires entre variables explicatives et variable expliquée. Certaines de ces liaisons n'auraient pas forcément pu être envisagées spontanément. En fort développement depuis trois décennies en lien avec le progrès technologique réduisant les temps de calcul et la prolifération des sources de données, elles s'avèrent souvent performantes mais l'interprétation des contributions des variables à la prédiction reste souvent malaisée.

En dépit du fait que ces techniques d'apprentissage statistique poursuivent des objectifs similaires aux approches plus classiques mentionnées précédemment, à savoir estimer une fonction de liaison à partir d'un nombre limité d'observations, ces approches ont été assez peu appliquées à l'analyse de survie. Ceci s'explique d'abord par les caractéristiques de ces données. En effet, l'évènement étudié (par exemple une rechute ou un décès) n'est pas observé chez tous les individus, certains patients peuvent ne jamais rechuter, ou bien rechuter en dehors de la fenêtre d'observation de l'étude ou encore être perdu de vue (parce qu'il déménage par exemple). Dans ce contexte, il peut être difficile de distinguer absence d'évènement et non observation d'un évènement qui a effectivement lieu. Les données sont donc incomplètes et dites censurées. La deuxième raison est davantage d'ordre méthodologique. Les approches de machine learning sont mises en place dans un unique but prédictif, fournissant un indicateur de qualité de la performance. Il s'agit donc de calibrer le modèle sur une partie des données disponibles et de tester sa capacité de généralisation à des observations *nouvelles*. Au contraire, les modèles statistiques classiques vont chercher à estimer le vrai modèle de génération des observations disponibles en faisant l'hypothèse que si le modèle est correctement identifié, alors les prédictions que l'on pourra éventuellement fondées sur ce dernier seront bonnes, mais ce n'est pas le but premier de l'estimation du modèle qui vise davantage à expliquer un phénomène. En particulier, dans le cas où le modèle serait mal spécifié, un bon ajustement aux données de l'échantillon ne garantirait toutefois pas une bonne qualité des prédictions.

La littérature est maigre concernant la modélisation de la probabilité de récidive pour les suicidants, mais ces méthodes ont bien été utilisées dans d'autres domaines faisant intervenir des données aux caractéristiques semblables. Ainsi, des techniques d'apprentissage statistiques ont été mobilisées pour la prédiction de la récidive criminelle comme alternative à la régression logistique et à l'analyse discriminante. Plus précisément, en analyse de survie, on peut trouver différentes façons de mettre en oeuvre les méthodes classiques d'apprentissage statistique. Le plus naturel consiste à convertir le problème d'analyse des durées en un problème de classification binaire. On peut par exemple, considérer le statut d'un patient à un instant donné. Si la durée considérée est faible par rapport à l'intervalle de suivi de l'étude, la censure liée à la fenêtre d'observation sera réduite. Toutefois la censure liée à la disparition d'un individu (déménagement, décès faisant suite à une récidive ou pour une autre raison) ne pourra être traitée. D'autres méthodes consistent à considérer une paire d'individus quelconques et à identifier le plus à risque des deux (Belle *et al.*, 2011). Seulement, les individus ne sont pas toujours comparables : ils le sont si la récidive de l'un a effectivement lieu avant la récidive ou censure de l'autre. Récemment, des techniques d'apprentissage ont été adaptées pour estimer la survie et non plus un état 0 ou 1. C'est notamment le cas d'une variante des forêts aléatoires.

Nous avons d'abord testé les techniques classiques de classification binaire pour la survenue d'une récidive (1) ou son absence (0), dans l'absolu ou sous 48h. Puis nous avons mis en oeuvre les forêts aléatoires pour la survie, afin de comparer ces deux approches. Après une brève description des données, nous présenterons succintement les différents algorithmes mobilisés ainsi que le protocole retenu pour gérer la rareté de l'évènement prédit ici, enfin nous présenterons les résultats préliminaires.

2 Données

L'objectif principal de ce papier consiste à évaluer et proposer un modèle probabiliste de récidive à court terme élaboré à partir des données enregistrées dans le cadre du Plan Régional de Santé Suicide, en Rhône-Alpes. Ces données ont été collectées dans le cadre d'une enquête organisée sur la période 2002-2007. En particulier ont été recueillies 13570 admissions aux urgences de 20 établissements hospitaliers pour tentative de suicide de 11239 patients distincts. Les récidives représentent donc 17 % des tentatives renseignées dans la base.

Les caractéristiques disponibles pour l'analyse sont les caractéristiques socio-démographiques du suicidant telles que l'âge au moment de l'hospitalisation, le sexe, la date et ville de naissance, code postal et lieu de résidence, la catégorie socio-professionnelle, l'état matrimonial, le montant et l'origine des ressources. D'autre part, on dispose pour chaque observation, d'informations relatives à la tentative de suicide : la date et l'heure du suicide, la présence de tentatives de suicide antérieures, l'année de la première tentative de suicide, l'année de la première hospitalisation de l'individu pour motif psychiatrique, la présence d'antécédents familiaux de tentatives de suicide ou encore le mode opératoire retenu par le suicidant. Des caractéristiques liées à la prise en charge sont également disponibles comme le nom et code du centre d'accueil, le service d'admission du suicidant, le type de prise en charge à l'entrée et la sortie du suicidant, la date et heure du contact avec l'intervenant médical, l'identifiant de cet intervenant qui ne sera pas utilisé ici, le nombre d'intervention avec le suicidant. Enfin, des caractéristiques cliniques sont également renseignées : l'affectation psychiatrique en évolution du suicidant, le score à l'échelle d'intentionnalité de Beck, les caractéristiques du traitement médicamenteux, des données non psychiatriques (affection, antécédents et addictions). Ces données déclaratives possèdent certaines faiblesses, la principale provenant de l'incomplétude des modalités disponibles pour la plupart des variables qualitatives. Par conséquent, il est en général impossible de distinguer les valeurs manquantes de la modalité *restante* : l'absence de modalité de la variable traitement peut être liée à l'absence de traitement ou à la non communication de cette information.

Le tableau 1 en annexe répertorie les fréquences de certaines modalités au sein de la population des récidivistes et non récidivistes (ou censurés). Cette première analyse suggère qu'il existe bien une différentiation au niveau de l'âge notamment, de l'intégration sociale du patient, de son traitement ou de son affectation psychiatrique.

3 Méthodologie

3.1 Algorithmes de classification

Dans une premier temps, le problème a été abordé comme un problème binaire, c'est-à-dire qu'on a cherché à prévoir s'il y aurait ou non récidive en fonction des caractéristiques individuelles de l'individu au moment de son hospitalisation pour tentative. Les algorithmes testés sont donc des algorithmes classiques de classification binaire comme le classifieur bayésien naïf, la régression logistique pénalisée ou encore les aggrégations d'arbre de classification (*bagging*, forêts aléatoires ou *boosting*).

Le classifieur bayésien naïf est un algorithme simple d'implémentation et peu consommateur de temps de calcul grâce à une hypothèse forte d'indépendance des covariables simplifiant grandement les calculs. Notons Y la variable prédite prenant les labels $m \in \{0, 1\}$, et $X \in \mathcal{M}_{n,k}$ les covariables. La formule de Bayes fournit la probabilité conditionnelle *a posteriori* d'être dans l'état m (formule de gauche) et l'hypothèse d'indépendance simplifie les calculs de cette formule (à droite) :

$$p(y = m^*|x) = \frac{p(x|y = m^*)p(y = m^*)}{\sum_{m=1}^{M} p(x|y = m)p(y = m)} \qquad p(x|y = m) = \prod_{j=1}^{k} p(x_j|y = m)$$

L'état prédit sera celui dont la probabilité sera maximale. Bien que l'hypothèse d'indépendance ne soit bien souvent pas très réaliste, ces classifieurs se sont avérés performants dans un certain nombre d'applications.

Compte-tenu du nombre assez important de modalités dans notre base (une centaine), il n'est pas invraisemblable que l'on se trouve en situation de multicolinéarité. Les régressions pénalisées sont utiles dans ce contexte car elles permettent d'obtenir des estimations plus robustes des coefficients en contrôlant leur amplitude et donc des prédictions moins sensibles à une faible variabilité d'une covariable. De plus, il est fort possible que seule une minorité des caractéristiques influent réellement sur la probabilité de récidive, il est donc raisonnable de penser que le modèle sous-jacent devrait être relativement parcimonieux (c'est-à dire avec une majorité de coefficients nuls). Dans un tel contexte, il est commun d'utilisée une pénalisation LASSO (Tibshirani, 1996) qui, en contraignant une partie des coefficients à s'annuler, procède à une sélection des variables les plus pertinentes simultanément à l'estimation du modèle. L'*Elastic Net* (Zou & Hastie, 2005), est une pénalisation plus générale, qui permet de retenir des variables importantes mais fortement corrélées, tout en procédant également à une sélection. L'influence de la pénalité est contrôlée par les paramètres α et λ, la formulation pour la régression linéaire est la suivante, mais la pénalisation s'applique de façon similaire à la régression logistique que nous utilisons ici :

$$\forall \lambda > 0 \ \hat{\beta}^{ElasticNet} = \arg\min_{\beta \in \mathbb{R}^k} ||Y - X\beta||_2^2 + \lambda \left[(1 - \alpha)|\beta|_1 + \alpha||\beta||_2^2 \right]$$

Les agrégations d'algorithmes ont également de bonnes vertues. Le *bagging* (Breiman *et al.*, 1984), les forêts aléatoires (Breiman, 2001) et le *boosting* (Schapire *et al.*, 1998) sont des mé-

thodes qui ont fait leur preuve. Sonvent fondées sur l'algorithme CART (*Classification algorithm and regression trees*), qui consiste à séparer les données de façon itérative en utilisant un critère d'hétérogénéité, leur popularité provient de ce qu'en agrégeant des centaines d'arbres, on parvient à réduire significativement la variance de la prédiction à l'origine de faible qualité. Plus précisément, le *bagging* consiste simplement à agréger des arbres construits sur des échantillons *bootstrap* tandis que les forêts aléatoires introduisent une part d'aléa en tirant $q < k$ variables à chaque séparation de chaque arbre parmi les k variables disponibles. Les arbres ainsi agrégés sont plus divers que dans l'approche *bagging*. Le *boosting* agrège les arbres au sein d'un processus adaptatif, en attribuant davantage d'importance aux individus mal classés d'une étape à l'autre du processus. Les méthodes fondées sur les arbres permettent également, indirectement, de sélectionner les variables dont on peut finalement mesurer l'importance dans le modèle, même si on ne peut en inférer une relation stricte de causalité. En analyse de la survie, une variante des forêts aléatoires a été développée en modifiant le critère de séparation des données lors de la construction des arbres, en en fournissant une estimation de la survie et non un état binaire en sortie (Ishwaran *et al.*, 2008) (on note qu'il existe aussi une adaptation de la pénalisation LASSO au modèle de Cox (Tibshirani, 1997) non étudiée ici).

3.2 Gestion des classes déséquilibrées

Dans de nombreux domaines tels que la détection de la fraude, la classification de textes, le domaine médical, il n'est pas rare que la classe de la population à laquelle on s'intéresse le plus soit faiblement représentée dans les données. Dans ce contexte, le risque d'obtenir un classifieur affectant la classe majoritaire à toutes les nouvelles observations est alors crédible. Plusieurs solutions ont été développées pour traiter ce type de problèmatique, parmi lesquelles : le rééquilibrage des classes en amont de l'estimation du modèle, le choix d'un algorithme peu sensible aux déséquilibres des classes ou l'utilisation de coûts de mauvaise classification différenciés selon les classes. Le rééchantillonnage a été ici mis en oeuvre pour les techniques de classification binaire sous plusieurs variantes : par exemple en retirant aléatoirement des observations de la classe majoritaire (*sous-échantillonnage*), en dupliquant à l'inverse aléatoirement des observations de la classe minoritaire (*sur-échantillonnage*) afin de la gonfler artificiellement. Il existe des variantes comme la technique SMOTE qui consiste à ajouter des observations synthétiques obtenues par interpolation (Chawla *et al.*, 2002), etc.

3.3 Procédure de validation des modèles

Dans la pratique, les modèles sont estimés et calibrés par validation croisée sur 3 plis en utilisant une partie des observations dont on dispose (75 %) et sa performance est ensuite évaluée sur les 25 % restant afin de s'assurer que l'on serait capable d'identifier la récidive d'un individu présentant des caractéristiques légèrement différentes de celles des individus présents dans la base. Pour calibrer et évaluer la qualité des modèles, on doit définir une métrique mesurant la capacité du modèle à prévoir correctement les deux classes. Pour cela on utilise la matrice dite *de confusion* qui correspond au tableau de contingence des classes de référence et des classes prédites par le modèle. A partir de cette matrice, il est possible de calculer les indicateurs classiques suivants : le rappel (R) et la précision (P) qui correspondent au taux de classification correcte de la classe d'intérêt (ou positifs) et à la précision de ce classifieur (i.e parmi les in-

dividus identifiés positifs, ceux qui auraient réellement récidivé). On peut également calculer la sensibilité et la spécifivité qui correspondent au taux de vrais positifs et au taux de vrais négatifs (les taux de classification correcte des deux classes présentes). Pour maximiser ces différents critères simultanément, on peut utiliser les critères classiques suivants : la F-mesure et la G-moyenne, qui sont particulièrement indiquées en présence de classes déséquilibrées. Si l'on souhaite donner plus de poids à une classe, ici la récidive, on peut calibrer le poids de R et de P dans la formulation de la $F-mesure$, par un paramètre α ($\alpha = 0,25$ dans ce papier).

$$G-moy. = \left(\prod_{m=1}^{M} R\right)^{\frac{1}{M}} \qquad F = \frac{2\times R\times P}{R+P} \qquad F_\alpha = \frac{1}{\frac{\alpha}{P}+\frac{1-\alpha}{R}} = \frac{(1+\beta^2)\times P\times R}{\beta^2\times P+R}$$

4 Résultats préliminaires

Les résultats des différentes simulations effectuées ont été représentés figure 3 en annexe en termes de rappel et de précision, que l'on cherche à maximiser simultanément. On peut voir qu'il est difficile de correctement prédire la récidive sans accepter une faible précision. Les meilleurs modèles au sens de la $F-0,25-mesure$ permettent d'identifier entre 60 et 70 % des récidivistes au prix d'une précision de 1/4 à 1/3. En termes de sensibilité et de spécificité, cela signifie qu'on arrive à prédire entre 60 et 70 % des récidivistes et environ la même proportion de non récidivistes, ce qui est bien entendu meilleur qu'un tirage aléatoire. Des taux légèrement plus élevés sont obtenus pour la prédiction de la récidive sous 48h, mais les classes étant encore plus désequilibrées, cela se traduit par une forte imprécision de l'algorithme. On peut constater que les différents algorithmes présentent des performances relativement comparables à condition d'avoir opéré un bon prétraitement sur les données, en particulier le réechantillonnage s'avère ici indispensable sans quoi, même en utilisant la $G-moyenne$, la $F-mesure$ ou la $F-0,25-mesure$, le rappel est très faible. L'utilisation des forêts aléatoire permet de mesurer l'importance des variables dans la construction du modèle, on a représenté cet indicateur figure 4 en annexe. Les variables les plus influentes semblent être le jour de semaine auquel la tentative a été faite, le nombre de tentatives précédentes, l'âge et le traitement suivi.

Les résultats obtenus par les forêts aléatoires adaptées à l'analyse de la survie sont très inégaux figures 1 et 2. On peut toutefois voir que pour certaines durées suffisamment élevées (6 mois), il est possible d'obtenir un rappel et une précision relativement proche de ceux des approches binaires, sans utiliser le moindre prétraitement.

5 Conclusion

Il semble difficile de prévoir plus de 50 % des récidivistes sans retenir par erreur 30 à 40 % de faux positifs, quelle que soit la méthode testée. Toutefois, il existe d'autres approches d'apprentissage statistique adaptées à la survie, inspirées du modèle de Cox ou non paramétriques (Hothorn *et al.*, 2006) ou, plus récemment des *support vector machines*, technique éprouvée pour la classification ou la régression. En outre, le calibrage de ces modèles peut également mobiliser des critères propres à l'analyse de survie (Mogensen *et al.*, 2012), (Heagerty *et al.*, 2000). Ces différentes pistes restent à approfondir.

Références

BELLE V. V., PELCKMANS K., HUFFEL S. V. & SUYKENS J. (2011). Support vector methods for survival analysis : a comparison between ranking and regression approaches. In *Journal Artificial Intelligence in Medicine archive*, volume 53, p. 107–118.

BREIMAN L. (2001). Random forests. In *Machine Learning*, volume 5.

BREIMAN L., FRIEDMAN J., OLSHEN R. & STONE C. (1984). Classification and Regression Trees. In *Wadsworth*.

CHAWLA N. V., BOWYER K. W., HALL L. O. & KEGELMEYER W. P. (2002). Smote : Synthetic minority over-sampling technique. In *Journal of Artificial Intelligence Research*, volume 16, p. 321–357.

COX D. & OAKES D. (1998). Analysis of Survival Data. In *Chapman and Hall, London*.

EFRON B., HASTIE T., JOHNSTONE I. & TIBSHIRANI R. (2004). Least angle regression. In *The Annals of Statistics*, volume 32, p. 407–499.

HASTIE T., TIBSHIRANI R. & FRIEDMAN J. (2001). The Elements of Statistical Learning ; Data mining, Inference and Prediction. In *Springer Verlag, New York*.

HEAGERTY P., LUMLEY T. & PEPE M. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. In *Biometrics*, volume 56, p. 337–344.

HOTHORN T., BUHLMANN P., DUDOIT S., MOLINARO A. & VAN DER LAAN M. (2006). Survival ensembles. In *Biostatistics*, volume 7.

HOTHORN T., LAUSEN B., BENNER A. & RADESPIEL-TROGER M. (2004). Bagging survival ensembles. In *Statistics in Medicine*, volume 23, p. 77–91.

ISHWARAN H., BLACKSTONE E., LAUER M. & POTHIER C. (2004). Relative risk forests for exercise heart rate recovery as a predictor of mortality. In *CJ. Amer. Stat. Assoc*, volume 99, p. 591–600.

ISHWARAN H., KOGALUR U., BLACKSTONE E. & LAUER M. (2008). Random survival forests. In *The Annals of Applied Statistics*, volume 2, p. 841–860.

KALBFLEISCH J. & PRENTICE R. (1980). The Statistical Analysis of Failure Time Data. In *Wiley, New York*.

KATTAN M., HESS K. & BECK J. (1998). Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. In *Computers and Biomedical Research*, volume 31, p. 363–373.

LEBLANC U. B. & CROWLEY J. (1992). Relative risk trees for censored survival data. In *Biometrics*, volume 48, p. 411–425.

LEBLANC U. B. & CROWLEY J. (1993). Survival trees by goodness of split. In *J. Amer. Stat. Assoc.*, volume 88, p. 457–467.

MOGENSEN U. B., ISHWARAN H. & GERDS T. (2012). Evaluating random forests for survival analysis using prediction error curves. In *Journal of Statistical Software*, volume 50.

SCHAPIRE R. E., FREUND Y., BARTLETT P. & LEE W. S. (1998). Boosting the margin : A new explanation for the effectiveness of voting methods. In *The Annals of Statistic*, volume 26, p. 1651–1686.

SEGAL M. (1988). Regression trees for censored data. In *Biometrics*, volume 44, p. 35–47.

TIBSHIRANI R. (1996). Regression shrinkage and selection via the lasso. In *J. Royal. Statist. Soc. B.*, volume 58, p. 267–288.

TIBSHIRANI R. (1997). The lasso method for variable selection in the cox model. In *Statistics in medecine*, volume 16, p. 385–395.

ZOU H. & HASTIE T. (2005). Regularization and variable selection via the elastic net. In *J. R. Statist. Soc.*, volume 67, p. 301–320.

A Analyse descriptive

Variables/modalités	**%**	**% parmi les non récid./les récid.**	**Signif. *p-value*** [1]
Sexe			
Femme	67.3	67.0 / 68.2	
Homme	32.7	33.0 / 31.8	
Âge			***
< 20 ans	12.8	14.2 / 9.2	
20-25 ans	11.2	12.3 / 8.4	
25-35 ans	19.9	19.5 / 21.0	
35-45 ans	25.4	23.4 / 30.5	
45-55 ans	19.0	17.8 / 22.2	
55-65 ans	7.6	7.9 / 6.9	
> 65 ans	4.0	4.9 / 1.8	
Mode de vie			***
Seul(e)	37.3	36.5 / 39.2	
Famille	49.6	51.2 / 45.5	
Non renseigné	13.1	12.3 / 15.3	
Affection Psychia.			***
Anxiété	6.7	7.3 / 5.1	
Dépression	39.9	40.3 / 38.9	
Troubles du comport.	1.0	0.9 / 1.0	
Handicap mental	3.5	2.8 / 5.2	
Troubles personnalité	7.1	6.2 / 9.4	
Autre diagnostic	3.1	3.3 / 2.4	
Non renseigné	38.8	39.1 / 37.9	
Traitement			***
Chimiothérapie	20.9	19.7 / 23.9	
Incitatins aux soins	15.2	15.0 / 15.8	
Aucun traitement	21.4	22.5 / 18.6	
Psychothérapie	16.6	17.6 / 14.0	
Autre traitement	5.3	4.8 / 6.4	
Non renseigné	20.7	20.5 / 21.4	
Récidivistes	**28.0%**		

TABLE 1 – Statistiques descriptives

B Résultats obtenus avec les forêts aléatoires pour l'analyse de survie

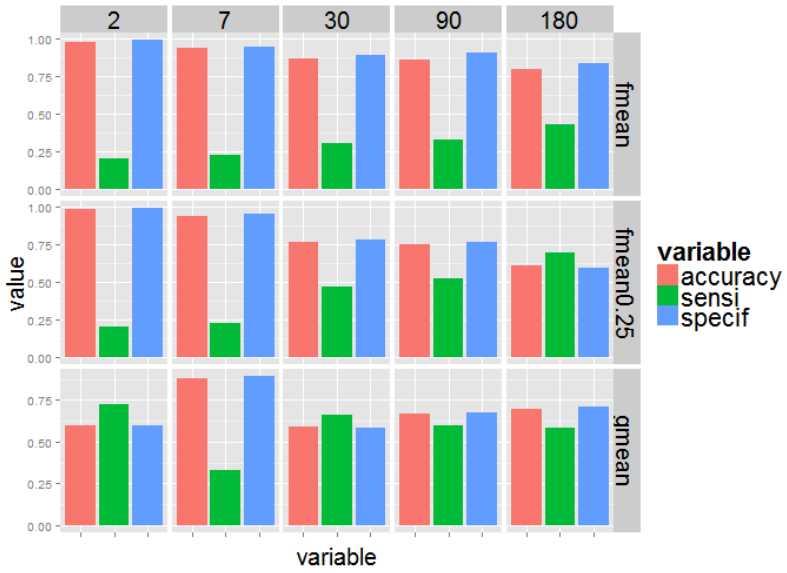


FIGURE 1 – Taux de vrais positifs et vrais négatifs pour les forêts aléatoires adaptées à l'analyse de survie à différentes durée (en jours)

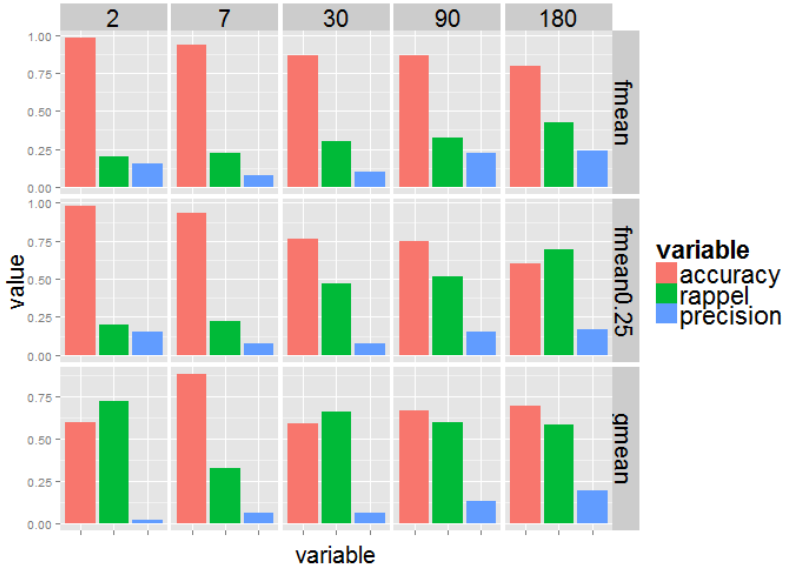


FIGURE 2 – Rappel et précision pour les forêts aléatoires adaptées à l'analyse de survie à différentes durée (en jours)

C Résultats obtenus avec les approches de classification binaire classiques

3.1 Résultats pour la détection de la récidive

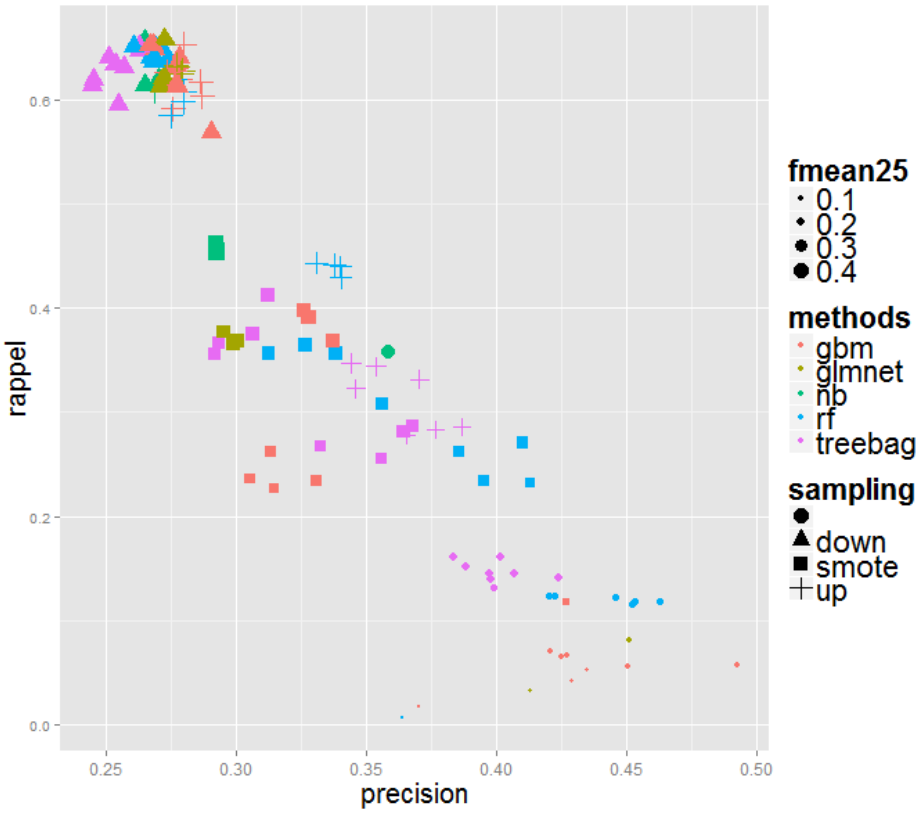


FIGURE 3 – Rappel et précision pour les différentes approches de classification binaire testées

Les résultats des différentes simulations effectuées ont été représentés en termes de rappel et de précision, que l'on cherche à maximiser simultanément. On peut voir qu'il est difficile de correctement prédire la récidive sans accepter une faible précision. Les meilleurs modèles au sens de la $F-0,25-mesure$ permettent d'identifier entre 60 et 70 % des récidivistes au prix d'une précision de 1/4 à 1/3.

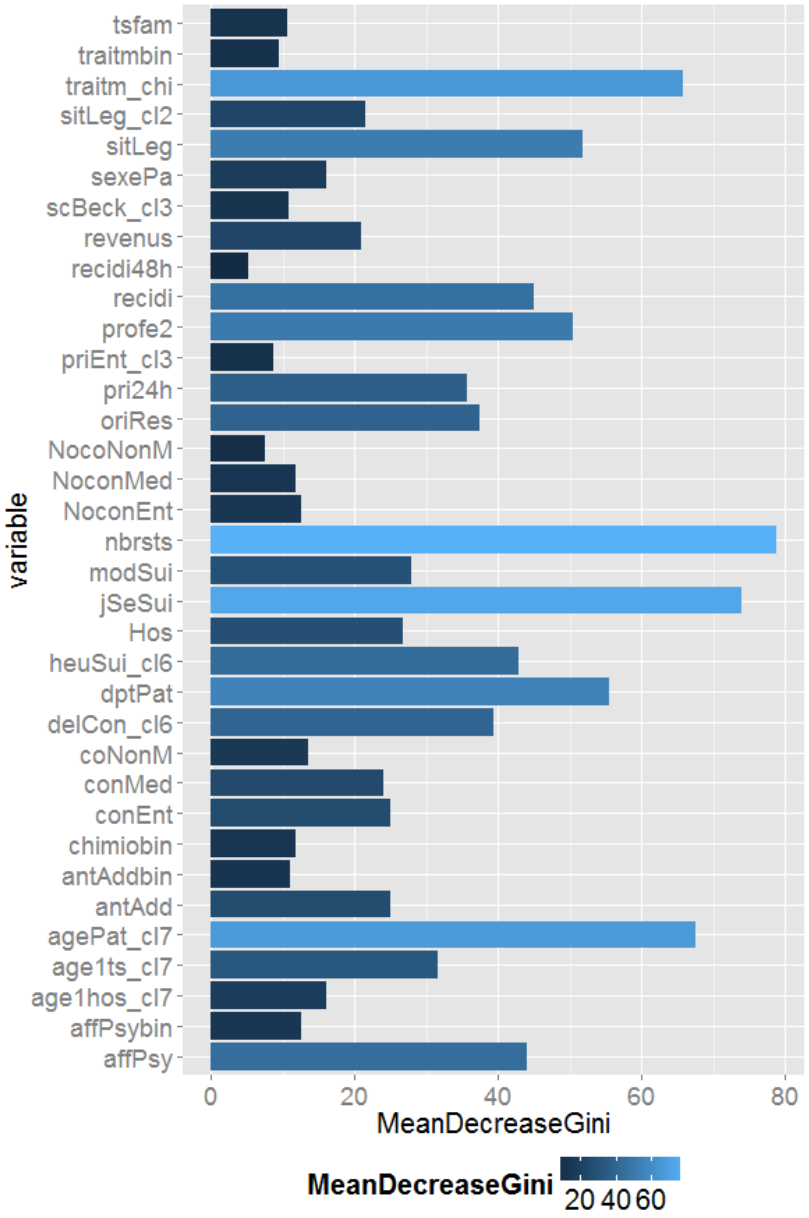


FIGURE 4 – Influence des variables dans un modèle utilisant les forêts aléatoires

Les variables les plus influentes au sens des forêts aléatoires semblent être le jour de semaine auquel la tentative a été faite, le nombre de tentatives précédentes, l'âge et le traitement suivi.

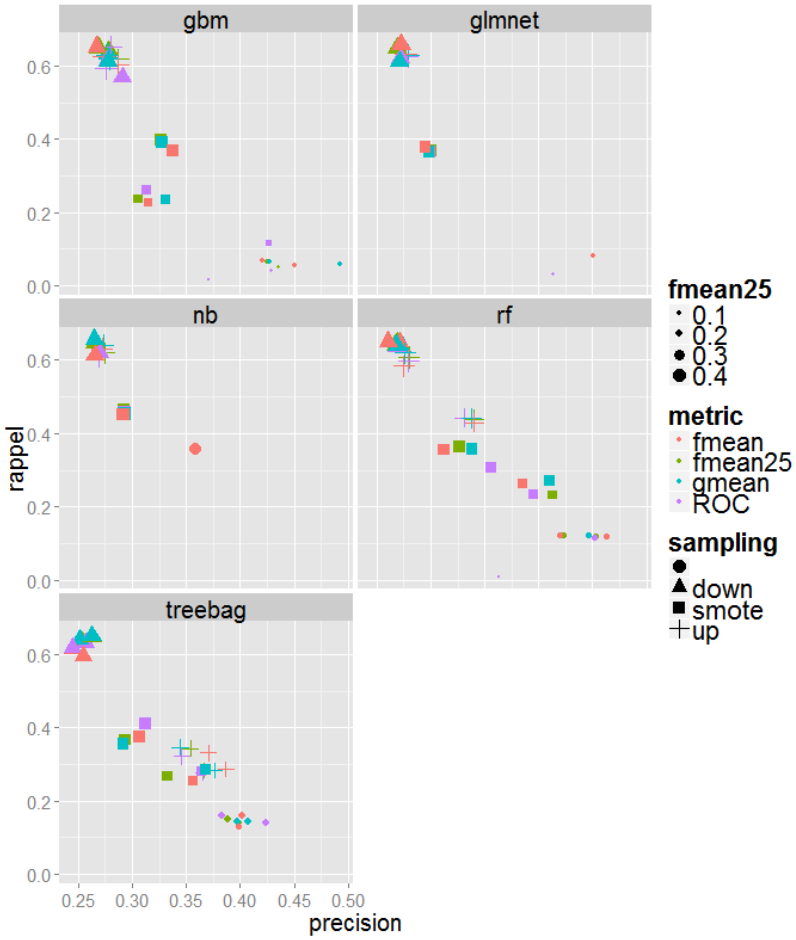


FIGURE 5 – Rappel et précision pour les approches classiques

Le réechantillonnage permet d'augmenter le rappel sensiblement (sous-échantillonnage et sur-échantillonnage en particulier). Les performances des différents algorithmes sur données rééchantillonnées sont comparables, le boosting se distingue du bagging mais les autres méthodes fournissent des résultats très proches.

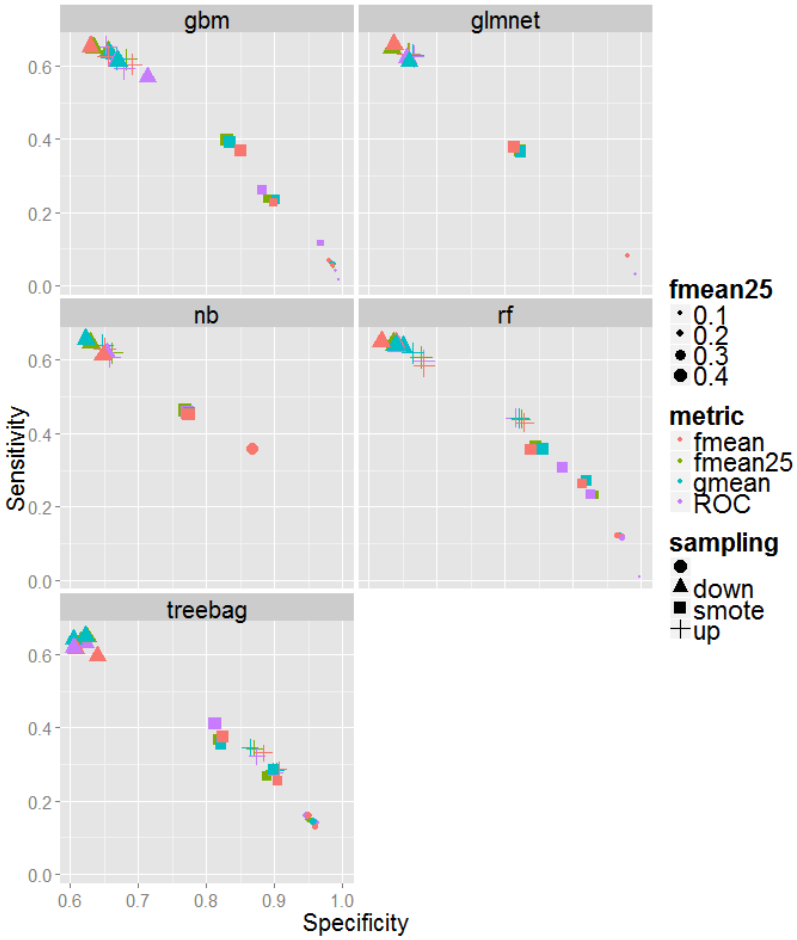


FIGURE 6 – Sensibilité et Spécificité (taux de vrais positifs et taux de vrais négatifs) pour les différentes approches testées

En termes de sensibilité et de spécificité, cela signifie qu'on arrive au mieux à prédire entre 60 et 70 % des récidivistes et environ la même proportion de non récidivistes, ce qui est bien entendu meilleur qu'un tirage aléatoire.

3.2 Résultats pour la détection de la récidive à court terme (48h)

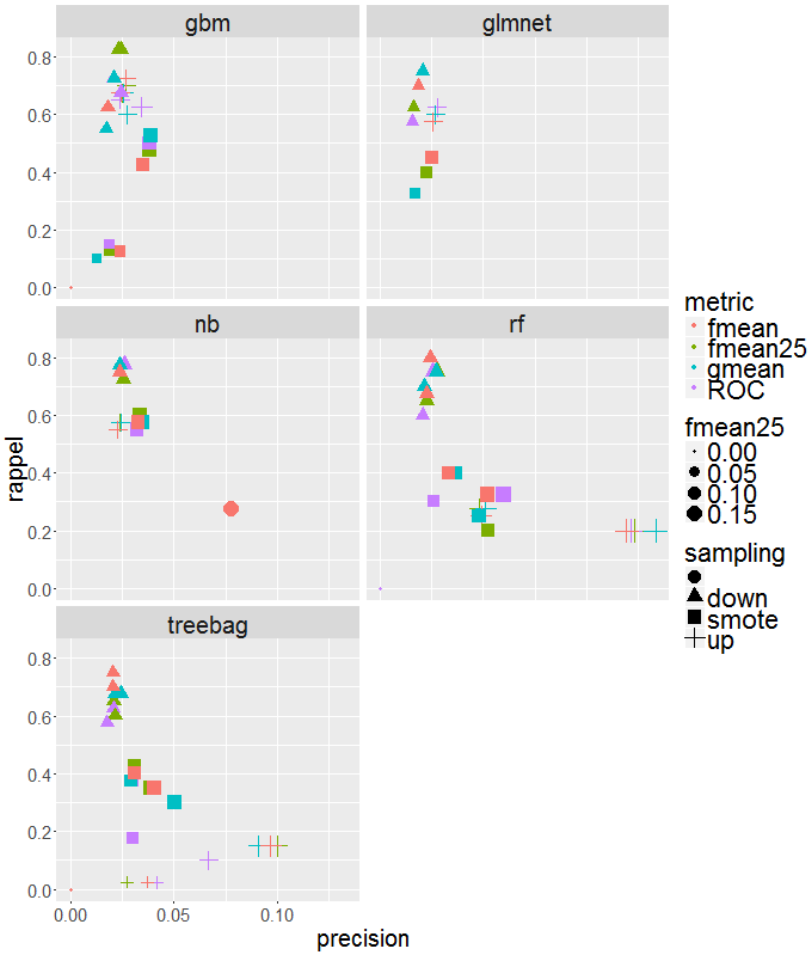


FIGURE 7 – Rappel et précision pour les approches classiques appliquées à la récidive à court terme

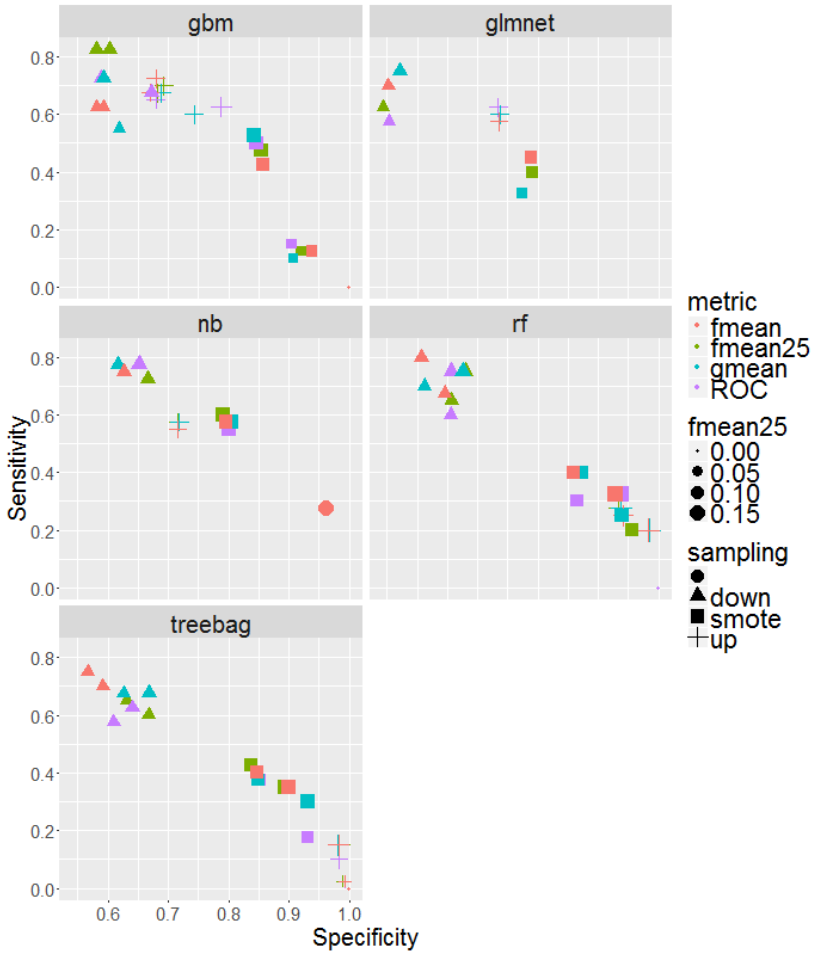


FIGURE 8 – Sensibilité et Spécificité (taux de vrais positifs et taux de vrais négatifs) pour les différentes approches testées

3.3 Résultats pour la détection de la récidive à une semaine

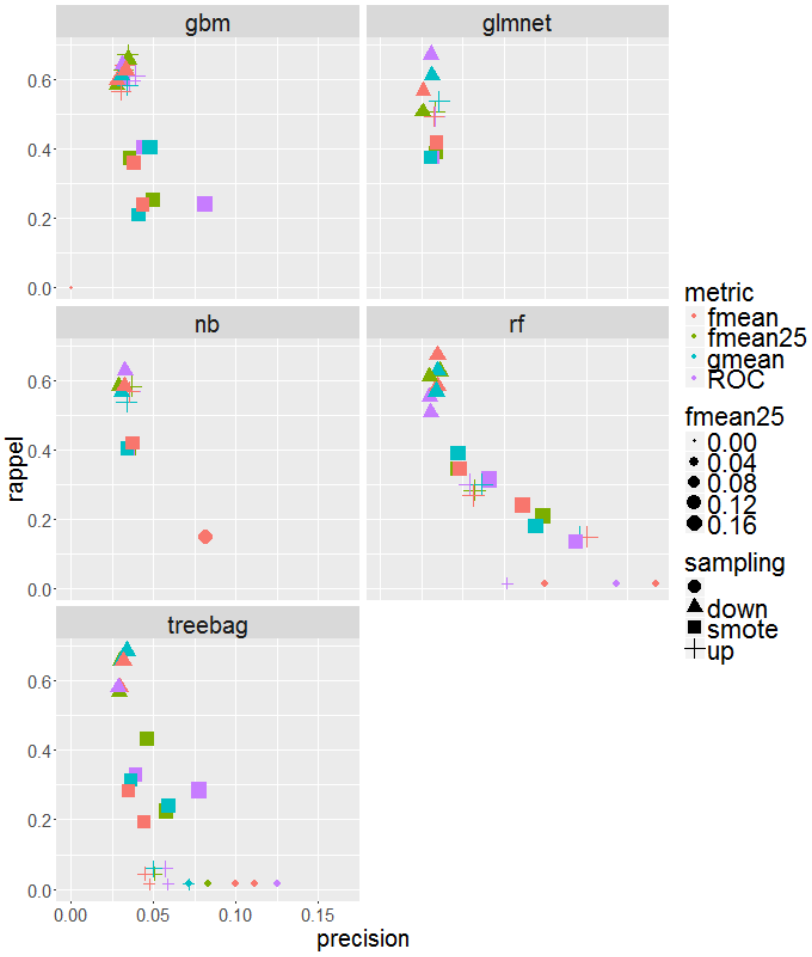


FIGURE 9 – Rappel et précision pour les approches classiques appliquées à la récidive à court terme

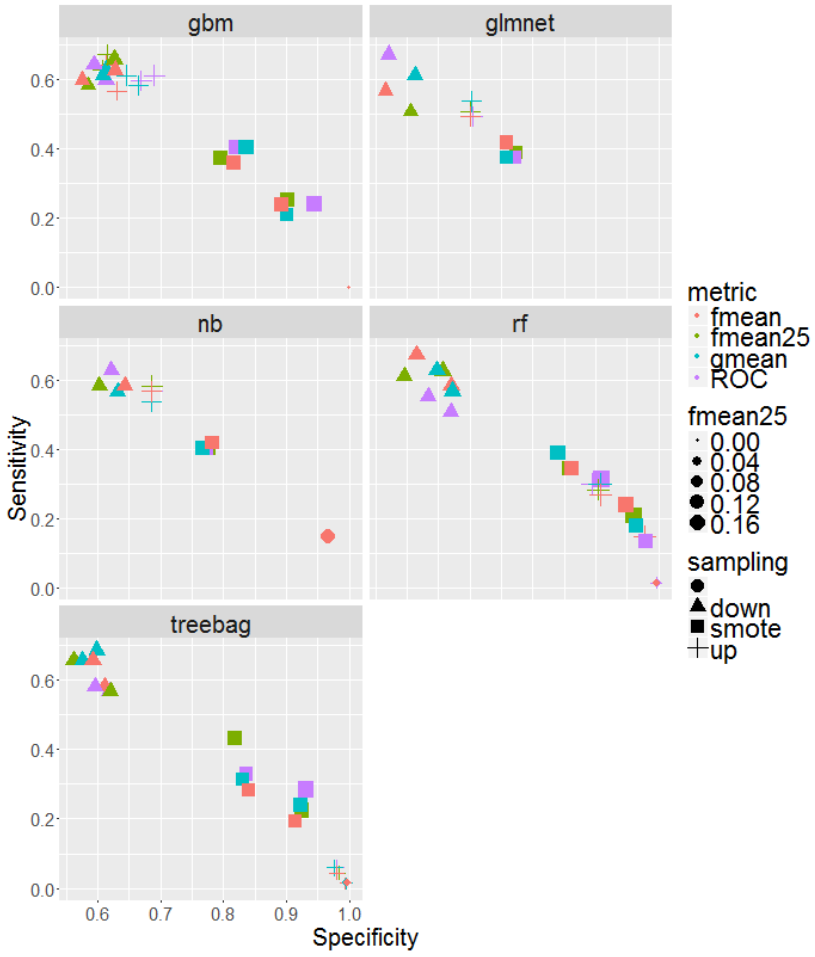


FIGURE 10 – Sensibilité et Spécificité (taux de vrais positifs et taux de vrais négatifs) pour les différentes approches testées

3.4 Résultats pour la détection de la récidive à 30 jours

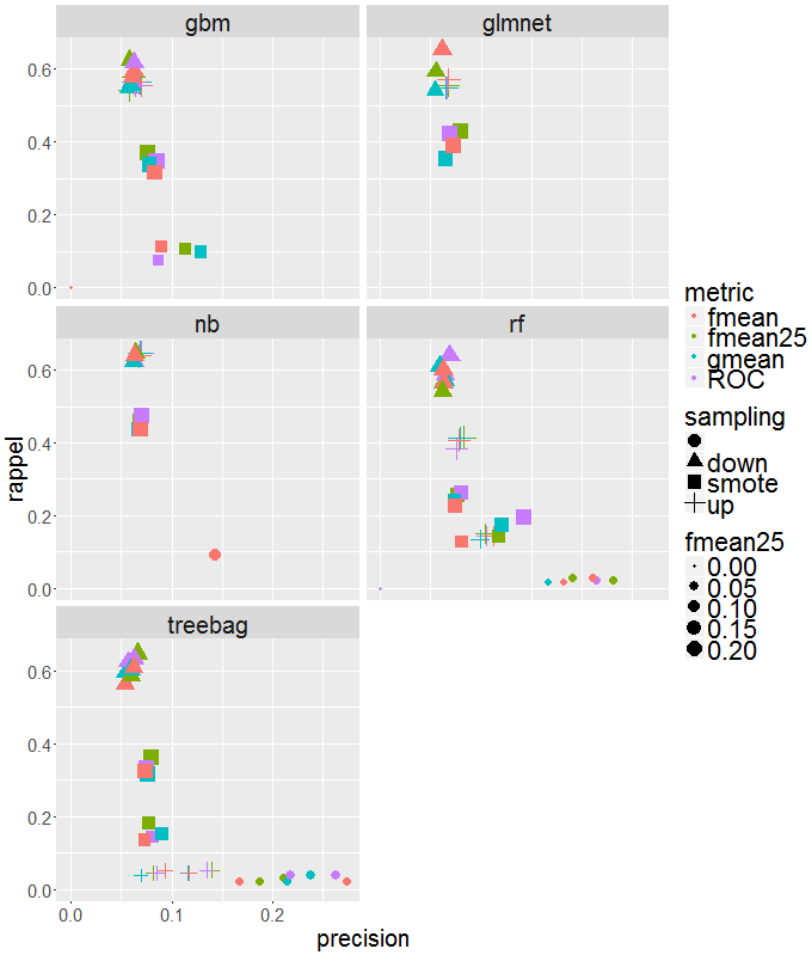


FIGURE 11 – Rappel et précision pour les approches classiques appliquées à la récidive à court terme

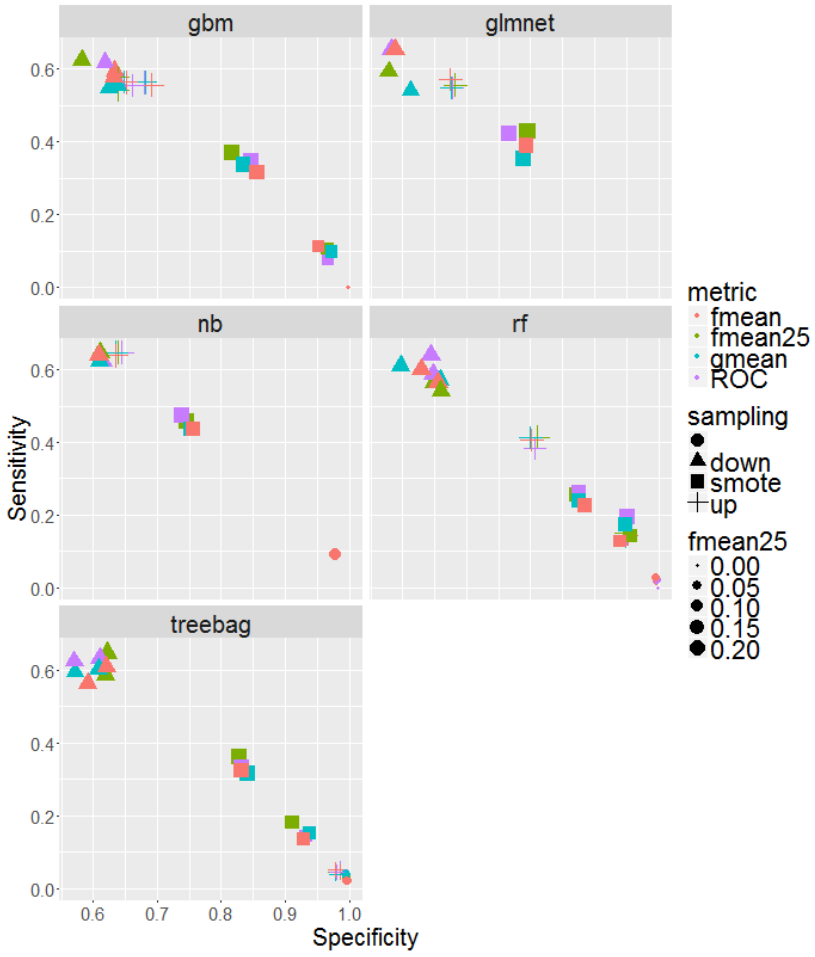


FIGURE 12 – Sensibilité et Spécificité (taux de vrais positifs et taux de vrais négatifs) pour les différentes approches testées

3.5 Résultats pour la détection de la récidive à 3 mois (90 jours)

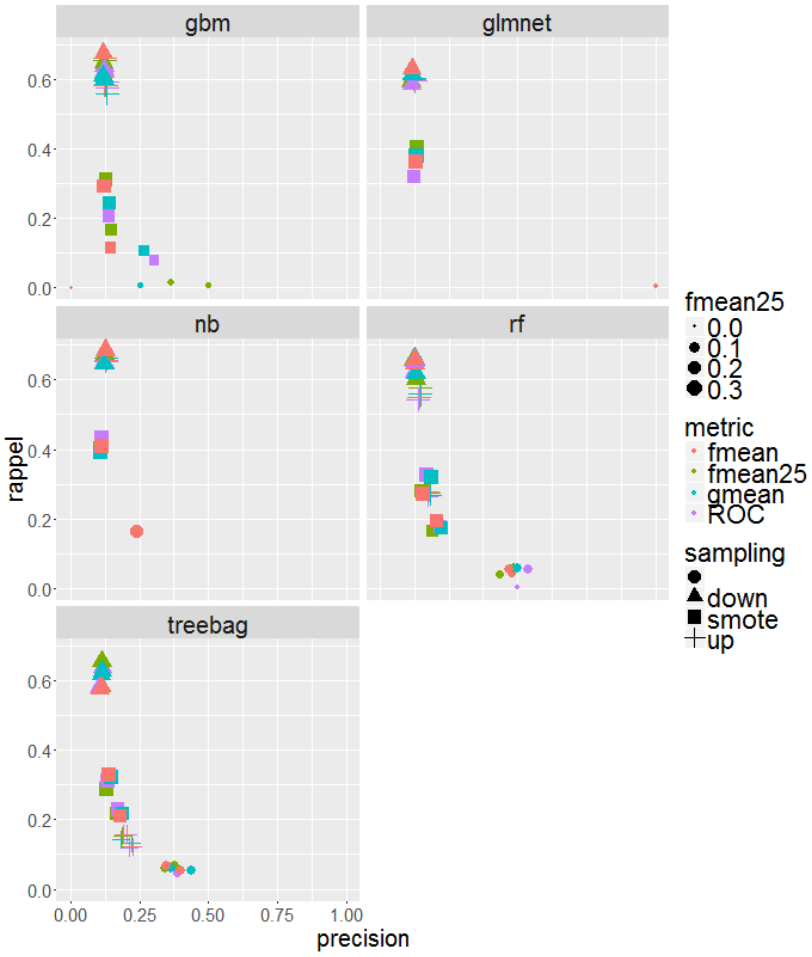


FIGURE 13 – Rappel et précision pour les approches classiques appliquées à la récidive à court terme

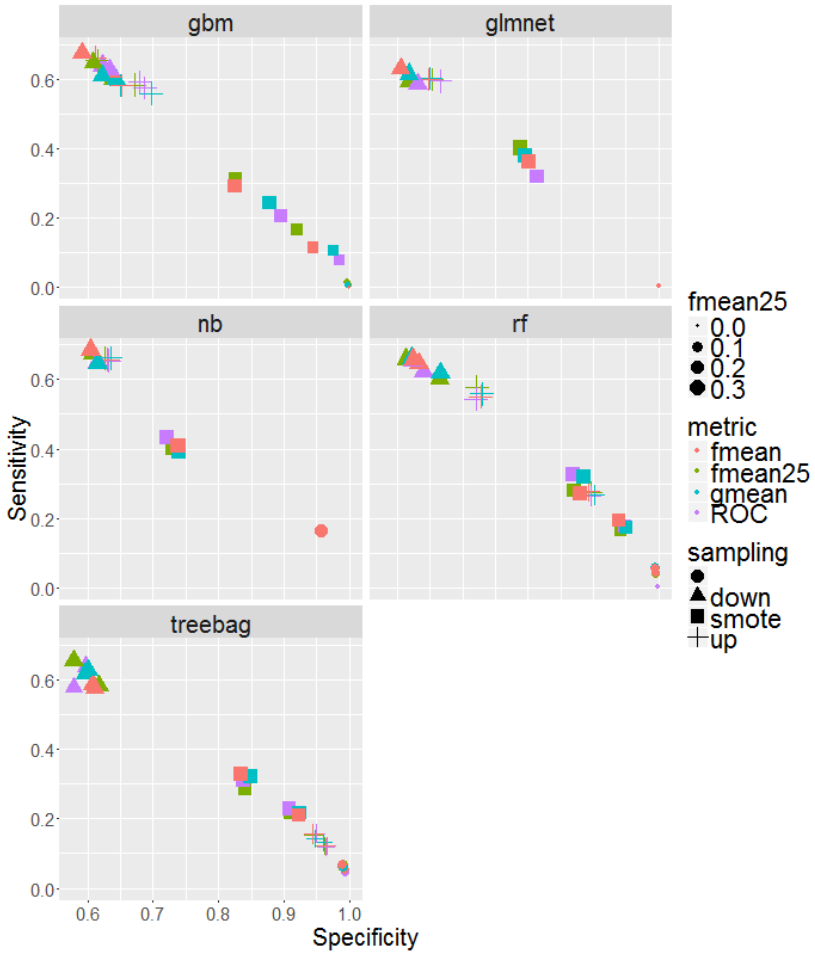


FIGURE 14 – Sensibilité et Spécificité (taux de vrais positifs et taux de vrais négatifs) pour les différentes approches testées

Les ontologies pour aider à comprendre les parcours de santé dans le cadre des maladies neurodégénératives[1]

Sonia Cardoso[1], Xavier Aimé[2], Luis Felipe Melo Mora[2], Marie-Christine Jaulent[2], David Grabli[3], Vincent Meininger[5], Jean Charlet[2,4]

[1] IHU-A-ICM Institut des Neurosciences Translationnelles de Paris,
s.cardoso-ihu@icm-institute.org
[2] INSERM UMRS 1142, LIMICS, F-75006, Paris
Sorbonne Universités, UPMC Univ. Paris 06, UMR_S 1142, LIMICS, F-75006, Paris
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse
xavier.aime@inserm.fr
luisfe.melo@gmail.com
[3] Assistance Publique Hôpital Pitié Salpêtrière, Département des maladies du Système Nerveux, Paris
Université Pierre et Marie Curie
david.grabli@psl.aphp.fr
[4]Assistance Publique –Hôpitaux de Paris DRCD, F-75004 PARIS
jean.charlet@upmc.fr
[5]Ramsay General de Santé, Hôpital Peupliers Paris
vincent.meininger@psl.aphp.fr

Résumé : Les maladies neurodégénératives touchent un grand nombre de personnes en France, elles se caractérisent par des répercussions multiples et polymorphes (médical, social et environnemental) altérant la qualité de vie des patients et de leurs entourages. En 2005, un réseau régional de coordination de parcours de santé de patients atteints de Sclérose Latérale Amyotrophique s'est mis en place, créant une base de données traçant les demandes, besoins des patients et les actions mises en place pour répondre aux besoins. Dans ce contexte, notre objectif est de construire, à partir des outils de l'ingénierie des connaissances, une ontologie. Nous utilisons pour cela, les ressources textuelles présentes dans les corpus, via le traitement automatique de la langue naturelle, afin d'exploiter cette base de données dans l'objectif de comprendre, décrire les parcours de santé et identifier les points de fragilité et les leviers pour améliorer la continuité des parcours et prévenir les ruptures.

Mots-clés : Ingénierie des connaissances, ontologie, parcours de soins, sclérose latérale amyotrophique.

1 Introduction

Les maladies neurodégénératives sont pour les politiques de santé une priorité nationale[2] et même internationale. Ces pathologies touchent un grand nombre de personnes : en France, 850 000 pour la maladies d'Alzheimer ou maladie apparentées, 150 000 pour la maladie de Parkinson et 85 000 pour la sclérose en plaques. « La prévalence de plus d'un million de

[1] Ce travail bénéficie du financement de l'IUIS (http://iuis.sorbonne-universites.fr/) dans le cadre du projet PARON

[2] En France elles bénéficient d'un plan spécifique « Plan Maladies neurodégénératives 2014-2019 » (PMND 2014). Cf. http://social-sante.gouv.fr/IMG/pdf/Plan_maladies_neuro_degeneratives_def.pdf

personnes en France et la gravité de l'impact de ces maladies sur la qualité de vie des personnes malades et de leurs aidants imposent une forte mobilisation » (PMND 2014). Or ces pathologies chroniques et évolutives ont un fort impact, sur la personne « malade », mais aussi sur les aidants naturels (l'entourage) et professionnels. Elles ont un coût financier direct et indirect conséquent au niveau sociétal. Les répercussions des pathologies – maladie d'Alzheimer et maladie apparentées, Sclérose latérale Amyotrophique (SLA) – sont multiples et polymorphes dans les atteintes et dans la temporalité, comme (1) l'altération de l'état de santé, (2) une diminution progressive et constante des capacités physiques et cognitives et (3) une altération de la qualité de vie. Elles ont également des conséquences sur les sphères sociales (interaction avec l'activité professionnelle, droit à la compensation afin de financer les aides humaines et techniques nécessaires à la réalisation d'activité de vie quotidienne) et environnementales. Le PMND met par conséquent en avant l'importance de travailler sur l'amélioration de la qualité de vie et de répondre aux attentes des personnes touchées.

Dans le cadre spécifique du suivi et de l'accompagnement des patients atteint de SLA, un réseau spécialisé en région Île-de-France, le réseau SLA IDF, a été créé en 2005 dont les objectifs sont (1) d'être un réseau ressource pour les patients, l'entourage et les professionnels de proximité intervenant dans la prise en charge des patients, et (2) de faire le lien entre la ville et l'hôpital et de façon spécifique avec le centre expert SLA de l'hôpital de la Pitié-Salpêtrière. Ces missions sont assurées par des coordonnateurs de parcours patient qui à travers une base de données tracent, par une saisie textuelle non structurée, l'ensemble des « évènements » intervenant dans le parcours de santé[3] des patients inclus dans le réseau (demandes, besoins, compte rendu de consultation, actions menées pour répondre aux besoins…). L'intervention régionale sur l'ensemble de l'Île-de-France du réseau SLA permet d'avoir une représentation globale des parcours de santé pour les patients puisque le réseau prend en charge 92% des patients atteints de SLA en Île-de-France (Cordesse *et al*., 2015).

La finalité de notre travail est de pouvoir comprendre et décrire les parcours de santé des patients atteints de cette pathologie neurodégénérative qui sert de modèle pour comprendre, identifier les points de fragilité et les leviers pour améliorer la continuité des parcours et prévenir les ruptures.

Pour atteindre cet objectif, nous utilisons les outils et méthodes de l'Ingénierie des Connaissances (IC) et du Traitement Automatique de la Langue Naturelle (TALN) afin de permettre une exploitation des données présentes dans les bases évènementielles et ainsi adopter une approche par le terrain « bottom-up ». Le premier travail consiste en la création d'une ontologie de domaine de la coordination des parcours de santé pour la SLA, afin de caractériser sémantiquement les événements de parcours. Gruber (1995) définit les ontologies comme une formalisation d'une conceptualisation partagée. Ce sont des systèmes informatiques de représentation formelle de la sémantique des concepts d'un univers de discours caractérisant un secteur d'activité. L'ontologie permet d'élaborer, de normaliser et de standardiser la sémantique associée à l'information tout en permettant à l'utilisateur d'employer une terminologie qui lui est familière pour décrire cette information. De nombreuses ontologies ont été réalisées dans le domaine de la médecine, *Ontolurgences* (Charlet *et al.*, 2012a), *Bilingual Ontology of Alhzeimer's disease and Related Diseases*, ONTOAD (Dramé *et al.,* 2014), et certaines spécifiques au domaine de la coordination des soins infirmiers, la *Nursing Care Coordination Ontology*[4], NCCO, mais aucune à ce jour ne regroupe les maladies neurodégénératives et la coordination de parcours de santé.

La première section de cet article présentera le projet et son contexte, la seconde montera la méthodologie utilisée et la dernière partie abordera les perspectives envisagées pour les maladies neurodégénératives.

---

[3] Article 1er de la Loi modernisation de notre système de santé modifiant article L. 1411-1 du code de la santé publique.

[...] La politique de santé comprend : [...] 5° L'organisation des parcours de santé. Ces parcours visent, par la coordination des acteurs sanitaires, sociaux et médico-sociaux, en lien avec les usagers et les collectivités territoriales, à garantir la continuité, l'accessibilité, la qualité, la sécurité et l'efficience de la prise en charge de la population, en tenant compte des spécificités géographiques, démographiques et saisonnières de chaque territoire, afin de concourir à l'équité territoriale [...].

[4] `http://purl.bioontology.org/ontology/NCCO`

2 Présentation du projet

2.1 La Sclérose Latérale Amyotrophique et le réseau SLA IDF

La SLA est une pathologie chronique neurodégénérative caractérisée par une dégénérescence progressive du motoneurone central et périphérique responsable d'un déficit moteur progressif. Ce déficit touche l'ensemble des muscles mais aussi le diaphragme et les muscles du pharynx et du larynx. Il engendre une perte d'autonomie progressive et rapide. Le pronostic vital peut être mis en jeu lors de l'atteinte respiratoire et des risques liés aux troubles de déglutition. Elle appartient au groupe des maladies rares, « son incidence en France est estimée à 2,5 pour 100 000 habitants et sa prévalence comprise entre 5 et 8 pour 100 000 habitants » (Couratier *et al*., 2014). Pour assurer au mieux le relai avec les dispositifs de prise en charge ambulatoire, il fut proposé, dés 2005, un modèle d'organisation structuré autour d'un réseau professionnel d'appui hospitalier tourné vers les soins ambulatoires : le réseau SLA IDF. Ce modèle intègre la notion de parcours (parcours de soins, parcours de santé, parcours de vie).

Afin d'assurer une traçabilité des actes, une base de données enregistrée à la CNIL fut créée. Y sont colligés les « événements » – c'est-à-dire les demandes émises par le patient son entourage ou les soignants quelle que soit la raison (médicale, sociale, environnementale) – ainsi que les actions de coordinations mises en œuvre pour répondre aux sollicitations. Ces événements résultent de la transcription écrite d'échanges oraux ou écrits, faits par les coordonnateurs, ainsi que la transcription d'éléments importants du parcours issus de documents plus formels comme les comptes rendus (compte rendu de consultation, compte rendu d'hospitalisation de jour en pneumologie, …). Ces données brutes, peu ou pas structurées, sont riches dans la compréhension de l'évolution naturelle de la pathologie. Elles reflètent aussi de façon réelle (concrète) et temporelle les situations complexes et critiques rencontrées à domicile par les patients et les professionnels de proximité, mettant parfois en péril la continuité des parcours de soins. A ce jour, la base contient 2 069 dossiers de patients et prés de 31 370 évènements.

2.2 Objectifs de modélisation

L'hypothèse de notre travail est qu'en se fondant sur une ontologie, il sera possible de réaliser une modélisation sémantique de la coordination et des événements permettant (1) une meilleure indexation des situations rapportées par les patients, (2) une optimisation de la recherche d'informations dans les bases de données et (3) des traitements statistiques notamment des facteurs prédictifs et la typologie des patients à risque. Pour cela, nous avons fait le choix de travailler spécifiquement sur la base de données du réseau SLA IDF. En effet, d'un point de vue médical, l'expression clinique de la pathologie par les différentes formes et atteintes (atteinte motrice, cognitive, respiratoire…) peuvent se retrouver dans de nombreuses autres pathologies neurodégénératives. De plus, ce réseau bénéficie d'une base de données conséquente en termes de durée d'existence, en nombre de dossiers ainsi qu'en diversité de situations de coordination.

Pour répondre à cet objectif, nous souhaitons pouvoir à travers les outils de l'IC modéliser des concepts issus de plusieurs domaines : médical (neurologie, pneumologie, nutrition…), social (handicap, allocation sociale, structure de proximité et étatique intervenant dans le parcours de santé), activités de coordination ayant une granularité assez fine pour exploiter l'ensemble des informations pertinentes présentes dans les corpus, sans être trop exhaustive. Ce niveau de granularité est défini conjointement par un travail de collaboration dans toutes les étapes de construction et d'annotation réalisée entre les ontologues et les experts de la neurologie et de la coordination.

Le choix des concepts de l'ontologie est fondamental. Ils doivent être représentatifs des aspects médicaux inhérents aux pathologies neurodégénératives. Ils doivent également être représentatifs des éléments de variation intervenant dans la mise en place des dispositifs de

soins et d'aide existants – qui ne sont pas linéaires et identiques dans toutes les situations – auprès des patients. Et ce, qu'il s'agisse (1) des choix de vie des patients (refus de soins, refus de mise en place d'aide humaine, demande pour l'organisation de séjour de répit, épuisement des aidants), (2) des disparités d'offre de soins et (3) parfois des disparités d'accès aux soins selon les situations géographiques (absence de médecin traitant, recherche de kinésithérapeute, mise en place de structure ambulatoire de type hospitalisation à domicile…).

La construction de l'ontologie et le traitement automatique des corpus nécessitent la mise en place de plusieurs outils, notamment par la diversité des corpus (structurés et non structurés) ainsi que la pluralité des formes orthographiques et syntaxiques (absence d'homogénéité) utilisés.

3 Méthodologie utilisée

3.1 Exploitation des corpus

Une première étape a consisté en l'extraction de termes du corpus à partir d'outils de TALN. Ce traitement s'est réalisé sur un corpus de 30 130 événements (structurés et non structurés) anonymisés extraits de la base SLA IDF, couvrant une période de dix ans d'activité du réseau (2005 à 2015). Ce corpus est large et abondant. Il est représentatif des problématiques médicales mais aussi des éléments médico-psycho-sociaux rencontrés par les patients au cours de la pathologie. Il répond aux exigences nécessaires pour couvrir le domaine. « Ce corpus doit être suffisamment large pour couvrir tout le domaine, et consensuel pour répondre à l'objectif d'une ontologie qui est – par définition – la formalisation d'une conceptualisation consensuelle » (Aimé, 2015)

Ce premier traitement c'est fait à partir du logiciel BIOTEX (Lossio-Ventura *et al*., 2014) permettant de choisir les candidats termes ayant un indicateur de fréquence important. Pour les candidats termes ayant un indicateur de fréquence moindre un travail de sélection s'est fait par les experts en neurologie.

3.2 Création de l'ontologie

A partir des syntagmes nominaux obtenus, un premier choix de classification pour la taxinomie fut réalisé permettant ainsi de catégoriser ces termes en quatre grands ensemble : (1) les agents, (2) les actions, (3) les états et (4) les objets. Ces quatre ensembles représentent les concepts clés permettant une classification et représentation optimale des éléments présents dans les corpus et représentatifs du contexte de coordination mettant en jeu des agents réalisant et intervenant sur des actions utilisant et modifiant l'état des objets.

Dans un premier temps, nous avons fait le choix de privilégier, pour construire l'ontologie, les termes issus de l'analyse des corpus, sans utiliser de classifications / terminologies comme par exemple la CIM-10, ou autres classification existante dans le domaine des aides techniques[5]. Une ontologie ayant pour objectif principal de représenter les connaissances d'un domaine suivant le point de vue d'un groupe d'individus, nous avons privilégié cette approche (bottom-up) afin d'obtenir l'ensemble des dénominations utilisées au quotidien par les agents et connaître les différentes formes orthographiques utilisées pour dans un second temps les lier aux classifications de référence. Un travail important a consisté en la recherche dans les corpus de tous les termes, acronymes et abréviations utilisés par les différents coordinateurs (parfois très nombreux) pour un même concept afin de les intégrer en tant que synonymes (*skos:altlabel*) dans l'ontologie et permettre ainsi, lors du travail d'annotation sémantique, un repérage des concepts.

[5] Telle que la Classification et terminologie des produits d'assistance pour personnes en situations de handicap-ISO 9999

Le travail de coordination implique la gestion simultanée de plusieurs situations ainsi que la réalisation de plusieurs tâches en même temps comme (1) les appels, (2) la réception et l'envoi de document et (3) la transcription dans la base des « évènements ». Ces conditions favorisent, afin de gagner du temps, l'utilisation d'un grand nombre d'abréviations et d'acronymes sans qu'il y ait pour autant d'homogénéité sur les termes utilisés. Par exemple, le concept *MédecinTraitant* peut être dénoté par les termes synonymes *médecin traitant*, *méd traitant*, *med ttt*, *MT*, *med tt*, ou encore *médecin de famille*. De même, pour désigner une structure intervenant au domicile du patient, il est utilisé le nom commercial de cette dernière et non son rôle, comme par exemple la structure X pour parler d'un prestataire délivrant les aides techniques. Une réunion de travail a donc été organisée afin de sensibiliser les coordinateurs sur l'importance d'unifier les termes choisis. Pour faciliter le traitement des corpus, une liste d'abréviations et d'acronymes fut donc choisi de manière consensuelle.

Un travail de recherche sur des ontologies existantes fut également réalisé afin d'identifier des concepts déjà existants en vue de les aligner. Les ontologies ONTOPYSCHIA (Richard *et al.*, 2013), ONTOLURGENCES (Charlet *et al.*, 2012a) et MENELAS (Charlet *et al.*, 2012ab) sont retenues de par leurs nombreux points de convergence avec notre travail. En effet, par la diversité des atteintes des maladies neurodégénératives, beaucoup de concepts présents dans les ontologies du domaine de la psychiatrie et de la médecine d'urgence sont utilisables pour l'ontologie de coordination de la SLA.

3.3 Modélisation intermédiaire des actes de coordination en UML

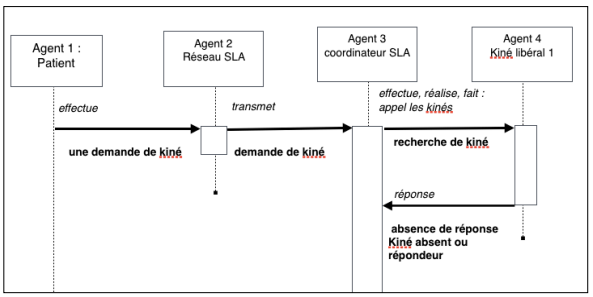


FIGURE 1- Extrait d'une modélisation d'action de coordination par les diagrammes de séquence.

Les actes de coordination peuvent être des actions simples comme la transmission d'une information ou bien plus complexes mettant en jeu de nombreux agents avec une demande qui peut évoluer et se modifier au cours de la réalisation de la tâche par l'intégration de nouveaux éléments. Afin de comprendre et modéliser de façon formelle les actions de coordination et intégrer les concepts dans l'ontologie, nous avons utilisé les diagrammes d'interaction comme le montre la figure 1. « Les diagrammes d'interaction ont pour but de décrire les modalités de communication entre les objets d'une application, d'un processus ou d'une organisation » (UML 2, 2006). La représentation par ce type de diagrammes a permis de distinguer pour de nombreuses actions de coordination les agents intervenants dans les processus, ainsi que la temporalité sous-jacente à ces processus. Dans cet exemple, la demande initiale du patient est de pouvoir bénéficier de soins de kinésithérapie, cette demande va se « transformer » en la recherche d'un kinésithérapeute par le coordinateur. La modélisation d'une action comme la demande avec un agent, un objet, un récepteur est évidemment liée aux travaux sur les *frames* de Minsky et les *scripts* de Schank[6] même si nous n'avons pas plus investigué ce sujet à ce jour. La notion de temporalité est également un élément important et multifactoriel, lorsque l'on travaille sur les parcours de soins. Il y a une temporalité médicale (évolution de la pathologie, avec parfois des situations d'urgence pouvant engager, à un stade avancé, le

[6] http://www.semantique-gdr.net/dico/index.php/Frames_et_Scripts

pronostic vital), une temporalité administrative (délais de traitement des dossiers, délais d'attente pour intégrer une structure de répit…). Cette succession d'activités et d'intégration des données est ordonnée et séquencée dans le temps, elle sera modélisée dans une étape future afin d'entrer dans le cadre de l'analyse du parcours.

3.4 Influence du travail d'annotation dans la construction de l'ontologie

L'ontologie vise à nous permettre de traiter et analyser de manière sémantique les données présentes dans les bases évènementielles et ce afin de mieux comprendre les parcours notamment par une représentation plus synthétique et homogène. En parallèle de l'élaboration de l'ontologie, un travail est mené sur la mise en place et le développement d'outils permettant l'annotation automatique des corpus. Ces travaux concomitants ont eu une influence réciproque riche et enrichissante, à la fois sur la construction de l'ontologie et sur la modularité des outils d'annotation. Le système d'annotation a en effet rapidement mis en avant les « lacunes » de l'ontologie (par exemple en permettant de faire le focus sur le manque d'exhaustivité parfois des termes utilisés par les coordinateurs et absents de l'ontologie, et non repéré comme candidats termes du fait de leur fréquence très faible). De plus, cette collaboration a permis de voir en temps réel, les erreurs des outils d'annotation permettant des actions d'amélioration rapides favorisant une optimisation des moyens et des résultats.

4 Etat d'avancement du projet

A ce jour l'ontologie est constituée de 2 480 concepts ; un travail important est encore à réaliser sur les définitions et la mise en place des relations, permettant le passage d'une simple taxonomie à une réelle ontologie. L'une des difficultés principales dans la création de l'ontologie est la modélisation des multiples actions, celles de coordination, les demandes des patients et professionnels intervenant dans le cadre des parcours. L'identification, la compréhension et la quantification de ces actions par le système d'annotation devrait donner des éléments de compréhension des parcours.

4 Conclusion et perspectives

Si actuellement le travail mené est réalisé sur la SLA, l'hypothèse de ce travail est qu'il sera transposable à d'autres pathologies neurodégénératives. Ce premier travail et l'ontologie crée ainsi que le système d'annotation sémantique devraient à terme être utilisable, avec certaines modifications dues aux spécificités de la maladie, à l'ensemble des bases de coordination dans le cadre des maladies neurodégénératives, permettant ainsi d'avoir une meilleure visibilité des parcours de soins de patients.

Nous prévoyons dans un proche avenir d'utiliser les données PMSI pour qualifier les parcours de soin dans la mesure ou les actes et les maladies déclarées dans ce contexte, peuvent être des marqueurs importants des parcours (Pinaire *et al.*, 2015).

L'association et la participation aux différentes étapes du projet des équipes de coordination et de neurologie, ont permis de montrer d'un point de vue du traitement informatique les difficultés engendrées par la diversité des abréviations et acronymes utilisés, ainsi que le manque parfois de structure syntaxique des phrases pour permettre une annotation complète. Il est possible de se questionner sur un devenir à court et moyen terme de l'impact prescriptif de ces difficultés sur le mode de saisies des évènements par les coordinateurs. Si l'analyse des bases va mettre en évidence les éléments des parcours de soins, par ricochet les activités de coordination vont également pouvoir être mises en lumière.

Références

AIME X. (2015) Eléments de réflexion sur l'utilisation de corpus pour la construction d'ontologies. IC 2015, Juin 2015, Rennes, France. AFIA, 2015.

CHARLET J., DECLERCK G., DHOMBRES F., GAYET P., MIROUX P.ET VANDENBUSSCHE P.-Y. Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In : Szulman S., coordinateur. Actes des 23es Journées Ingénierie des Connaissances, Paris, France, 27-29 juin 2012, p. 33-48.

CHARLET J., BACHIMONT B., MAZUEL L., DHOMBRES F., JAULENT M. ET BOUAUD J. OntoMenelas : motivation et retour d'expérience sur l'élaboration d'une ontologie noyau de la médecine. Technique et Science Informatiques, 31(1), 2012.

CORDESSE V., SIDOROCK F., SCHIMMEL P., HOLSTEIN J., MEININGER V. (2015) Coordinated care affects hospitlization and prognosis in amyotrophic lateral sclerosis : a cohort study. BMC Health Services Research

COURATIER P., MARIN B., LAUTRETTE G., NICOL M., & PREUX P.-M. (2014) Epidémiologie, spectre clinique de la SLA et diagnostics différentiels. La Presse Médicale, Volume 43, 538-548.

DRAME K., DIALLO G., DELVA F., DARTIGUES J-F., MOUILLET E., SALAMON R., MOUGIN F. (2014) Reuse of termino-ontological ressources and text corpora for building a multilingual domain ontology : An application to alzheimer's disease. Journal of Biomedical Informatics 48, 171-182

GRUBER T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. International journal of human-computer studies, 43(5), 907–928.

RICHARD M., AIME X., KREBS M.-O. & CHARLET J. (2013) Au delà du DSM : les ontologies comme aide aux classifications descriptives psychiatriques ? 2e édition du Symposium sur l'Ingénierie de l'Information Médicale, Jul 2013, Lille, France.

PILONE D., PITMAN N. (2006) UML 2 en concentré. Edition O'Reilly, Paris.

PINAIRE J., RABATEL J., AZE J., BRINGAY S., LANDAIS P. (2015) Recherche et visualisation de trajectoires dans les parcours de soins des patients ayant eu un infarctus du myocarde. 3ème Symposium Ingénierie de l'Information Médicale (SIIM 2015), Jun 2015, Rennes, France.

LOSSIO-VENTURA J.-A., JONQUET C., ROCHE M., & TEISSEIRE M. (2014) BioTex: A system for Biomedical Terminology Extraction, Ranking, and Validation – ISWC 2014 – 13th International Semantic Web Conference. Riva del Garda, Italy, 2014

Reconnaissance des stades de sommeil à l'aide d'un outil de support à la décision basé sur les connaissances et la pratique des experts

Adrien Ugon[1], Amina Kotti[1], Karima Sedki[2], Carole Philippe[3], Brigitte Séroussi[2,3], Jacques Bouaud[2,4], Jean-Gabriel Ganascia[1], Patrick Garda[1], Andrea Pinna[1]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, Paris, France
prenom.nom@lip6.fr

[2] Sorbonne Universités, UPMC Univ Paris 06, INSERM
Sorbonne Paris Cité, Université Paris 13, LIMICS, UMR_S 1142, Paris, France
[3] AP-HP, Hôpital Tenon, Département de Santé Publique, Paris, France
brigitte.seroussi@aphp.fr

[4] AP-HP, DRCD, Paris, France
jacques.bouaud@aphp.fr

[5] AP-HP, Hôpital Pitié-Salpêtrière, Unité Pathologies du sommeil, Paris, France
carole.philippe@aphp.fr

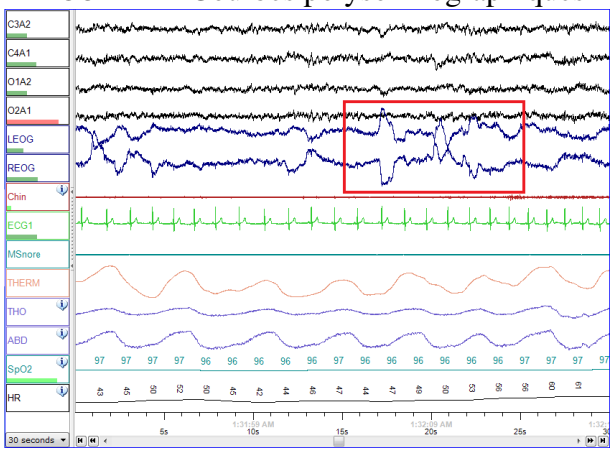
Résumé : La reconnaissance des stades de sommeil est une étape indispensable au diagnostic des troubles du sommeil. L'approche habituelle est d'utiliser un classifieur avec des méthodes d'apprentissage automatique. Nous proposons une approche innovante conçue à partir de la connaissance et de l'observation de la pratique des experts. Notre approche permet de mieux prendre en compte les aspects dynamiques du sommeil, mais aussi les raisonnements à des niveaux d'abstraction différents, ainsi qu'à des échelles de temps différentes. Après avoir extrait, à partir des signaux acquis, les *grapho-éléments* nécessaires à la décision en appliquant la fusion symbolique, un système expert incluant des règles d'inférence, intégrant si nécessaire des préférences, est appliqué. Mise en œuvre sur trois *époques* généralement mal scorées avec les classifieurs habituels, la méthode s'est avérée efficace.

Mots-clés : Système expert, Classification automatique, Outil d'aide à la décision, Stadification automatique du sommeil, Polysomnographie

1 Introduction

Le Syndrome d'Apnées du Sommeil (SAS) est un trouble du sommeil caractérisé par la survenue de pauses respiratoires fréquentes au cours du sommeil. Sa prévalence est estimée entre 3 et 7% (Marin *et al.*, 2005). Il est associé à une morbidité et une mortalité cardio- et cérébrovasculaires accrues (Punjabi, 2008; Arzt *et al.*, 2005), ce qui en fait un problème de santé publique. Il est également reconnu que les patients souffrant de SAS ont une dégradation de la qualité de vie (D'Ambrosio *et al.*, 1999), et qu'ils présentent un risque plus important d'accident de la circulation (Terán-Santos *et al.*, 1999) ou d'accident du travail (Uehli *et al.*, 2014).
La polysomnographie est l'examen standard pour le diagnostic du SAS. Il consiste en l'enregistrement simultané de différents paramètres physiologiques au cours d'une nuit complète. Les courbes obtenues (cf figure 1) sont alors analysées visuellement par des experts du sommeil afin de déterminer les stades de sommeil et les événements pathologiques.

FIGURE 1 – Courbes polysomnographiques



Actuellement, le scorage des courbes polysomnographiques est effectué visuellement par des experts du sommeil. C'est une tâche pénible et longue associée à une variabilité inter- et intra-scoreurs non négligeable (Kuna *et al.*, 2013). Il est reconnu que cette étape peut être améliorée (Penzel *et al.*, 2013).

La reconnaissance automatique des stades de sommeil est souvent considérée comme un problème de classification que l'on peut résoudre par apprentissage automatique. Des paramètres, extraits automatiquement par des méthodes de traitement du signal, sont ensuite injectés dans un classifieur qui indique en sortie l'un des cinq stades de sommeil définis dans les recommandations internationales (Alvarez-Estevez & Moret-Bonillo, 2015).

Aujourd'hui, les algorithmes de classification automatique inclus dans les logiciels hospitaliers ne satisfont pas les experts du sommeil (Escourrou *et al.*, 2010) ; l'analyse visuelle, bien que fastidieuse, est toujours préférée. Les critiques portent en premier lieu sur la fréquence des changements de stade de sommeil, jugée trop importante. D'après nous, cela est dû au traitement indépendant de chaque *époque*, à l'aveugle de ce qui a été décidé lors de l'*époque* précédente et de ce qui sera décidé à l'*époque* suivante. L'observation des pratiques de scorage dans l'unité de pathologies du sommeil de la pitié-salpêtrière (AP-HP) a mis en évidence que les médecins cherchaient davantage les ruptures dans le signal — analysées comme des transitions — que l'observation systématique d'éléments en faveur, ou en défaveur, d'un stade de sommeil. Le support des experts a permis une compréhension plus fidèle de leur pratique des recommandations internationales.

Nous avons choisi d'aborder le problème différemment en intégrant des raisonnements à des niveaux d'abstraction différents et à des échelles de temps différentes, permettant une meilleure prise en compte de l'aspect dynamique du sommeil. La succession des stades de sommeil n'est pas aléatoire. Notre système est basé sur un système expert. La base de faits est initialisée par fusion symbolique. Les stratégies de fusion et les règles du système expert sont inspirées de celles données par les recommandations internationales (Berry *et al.*, 2015).

Dans la prochaine section, nous verrons comment sont définis les paramètres du système expert. Puis nous testerons l'approche sur une portion de signal réel. Suivront une discussion puis une conclusion.

2 Méthode

2.1 Sources de connaissance

Le projet AEP (*Automatic Embedded Polysomnography*) a pour objectif d'améliorer la polysomnographie en développant un réseau de capteurs sans fils intégrant une intelligence capable d'interpréter les signaux. Nous nous intéressons ici aux traitements automatiques permettant la reconnaissance automatique des stades de sommeil. Dans ce projet, nous avons choisi de nous inspirer de la démarche cognitive des médecins interprétant les signaux pour créer l'outil de support à la décision. Pour cela, deux sources étaient à notre disposition : (1) les recommandations internationales en médecine du sommeil (2) l'observation de la pratique du médecin expert.
Des recommandations internationales pour le scorage des courbes polysomnographiques ont été définies par l'*American Academy of Sleep Medicine* (AASM) et sont publiées dans un manuel régulièrement mis à jour (Berry *et al.*, 2015). Cinq stades de sommeil sont définis : W (Eveil), N1, N2 (Sommeil lent léger), N3 (Sommeil lent profond) et R (Sommeil paradoxal). L'enregistrement complet, sur une nuit, est divisé en segments de 30 secondes, appelées *époques*. La stadification du sommeil consiste à assigner à chaque *époque* un des cinq stades de sommeil précités. Pour cela, l'expert doit s'appuyer sur des observations faites sur les trois courbes électro-encéphalographiques (EEG) mesurant l'activité cérébrale, les deux courbes électro-oculographiques (EOG) mesurant les mouvements oculaires et la courbe électromyographique (EMG) mesurant le tonus musculaire submentonnier. Les observations permettant de prendre la décision portent sur le rythme des ondes enregistrées et sur la présence de motifs spécifiques, appelés « *grapho-éléments* », des différents stades de sommeil.
Afin d'interpréter correctement les guides de bonne pratique, nous avons observé des médecins experts de l'unité de pathologies du sommeil à la pitié-salpêtrière (AP-HP) lors de l'interprétation des signaux polysomnographiques, et, en particulier, dans la stadification du sommeil.

2.2 Catégorisation des règles

En analysant le contenu du manuel de l'AASM (Berry *et al.*, 2015), nous avons divisé les règles en trois catégories :

— les définitions des *grapho-éléments* ;
— les règles d'inférence permettant de décider de l'observation locale d'un stade de sommeil ;
— les règles statistiques permettant d'assigner un stade de sommeil à l'échelle de l'*époque* en fonction des observations locales d'un ou plusieurs stades de sommeil. Elles sont basées sur la proportion de chaque stade de sommeil observé.

Certaines sont exprimées sous la forme de priorités. Cela concerne à la fois l'observation des *grapho-éléments* que l'observation de stades de sommeil. Lorsque plusieurs grapho-élements (resp. stades de sommeil) sont observés sur la même portion de signal, on donnera priorité à un grapho-élément (resp. stade de sommeil).

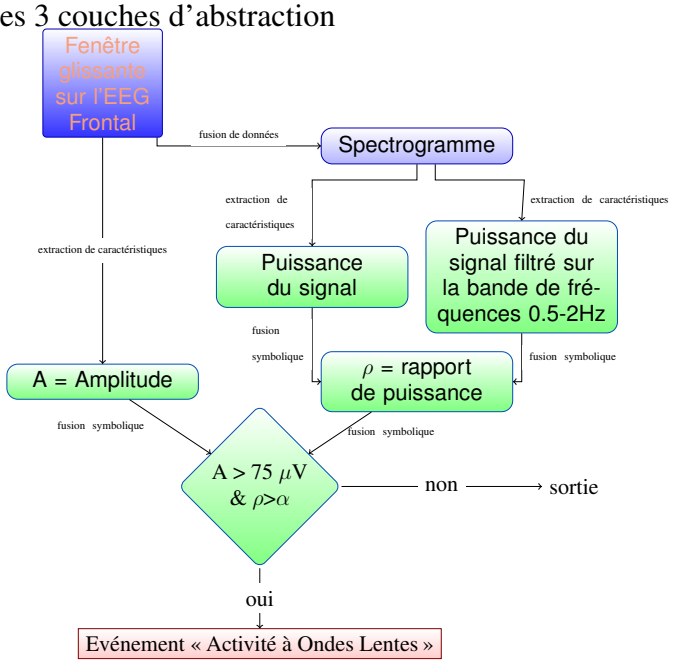
Les *grapho-éléments* sont extraits par fusion symbolique. Les observations locales des différents stades de sommeil sont décidées par un système expert ; les règles de priorité sont formalisées par des préférences et sont utilisées pour résoudre les conflits. Les règle statistiques sont alors simples à modéliser. Au total, les règles du manuel de l'AASM de scorage du sommeil de l'adulte ont été formalisées en 23 règles d'inférence dans notre système expert. Ces règles n'ont pas évolué dans les dernières mises à jour des recommandations, elles sont aujourd'hui consensuelles. L'actualisation des règles du système expert ne requerra donc, a priori, qu'une modification marginale par un technicien.

2.3 Formalisation des règles

2.3.1 Extraction des grapho-éléments

L'extraction des grapho-éléments se fait par fusion symbolique dans une architecture multi-couches (Ugon *et al.*, 2011). Cette architecture, définie par B. Dasarathy dans (Dasarathy, 1997), est constituée d'une couche contenant les données, d'une couche contenant les caractéristiques et d'une couche contenant les décisions. Des opérations de fusion et d'extraction permettent d'abstraire les données et les caractéristiques d'un niveau d'abstraction vers un autre.
À titre d'exemple, intéressons-nous à la reconnaissance de « l'activité à ondes lentes », *grapho-élément* spécifique du sommeil lent profond (N3). Selon l'AASM, par définition, il s'agit d'« *ondes de fréquence 0.5 à 2.0Hz avec une amplitude minimale de 75μV* ». L'extraction de cet événement suit le schéma de la figure 2. En bleu, on peut voir la couche de données, en vert, la couche des caractéristiques et en rouge la couche de décisions.

FIGURE 2 – Schéma d'extraction de l'événement « activité à ondes lentes » à partir de l'EEG frontal au travers des 3 couches d'abstraction



Les observations obtenues à cette étape servent de *base de faits* pour notre système expert.

2.3.2 Règles d'inférence du système expert

L'enregistrement complet est divisé en segments continus de courte durée (1 seconde). Les règles d'inférence permettent de décider, en fonction des « *grapho-éléments* » observés sur les segments courant, précédent et suivant, si un stade de sommeil est observé localement ou non.
À titre d'exemple, la règle G.2 du manuel de l'AASM nous indique de « *commencer à scorer le stade N2 (en l'absence des critères de N3) si l'une au moins des observations suivantes est faite : observation d'un ou plusieurs fuseau(x) du sommeil ; observation d'un ou plusieurs complexe(s) K non associé à un micro-éveil* ». Il est précisé dans la définition G.1 du manuel de l'AASM que « *pour qu'un micro-éveil soit associé à un complexe K, il faut que ce micro-éveil survienne concomitamment ou au plus tard une seconde après la fin du complexe K* ». Ces deux règles peuvent se formaliser par la formule 1.

$$\neg \mathscr{O}_{i-1}^{N2} \wedge \neg \mathscr{O}_{i}^{N3} \wedge \mathscr{O}_{i}^{\text{K complex}} \wedge \neg \left(\mathscr{O}_{i}^{\text{arousal}} \vee \mathscr{O}_{i+1}^{\text{arousal}} \right) \Rightarrow \mathscr{O}_{i}^{N2} \tag{1}$$

où :
— $\mathscr{O}_i^E$ signifie que l'événement E est observée sur le segment i (segment courant).
— $\mathscr{O}_i^S, S \in \{W, R, N1, N2, N3\}$ signifie que le stade de sommeil S est observé sur le segment i (segment courant).

2.3.3 Règles de priorité

Certaines règles définissent des priorités à respecter lorsque plusieurs stades de sommeil sont observés au cours d'une même *époque*. Nous les avons formalisées à l'aide de règles d'inférence intégrant des préférences.
À titre d'exemple, la règle I.4 du manuel de l'AASM indique que « *le stade R a priorité sur le stade N2* », ce qu'on peut formaliser avec la formule 2 :

$$\mathscr{O}_i^{N2} \wedge \mathscr{O}_i^R \Rightarrow \mathscr{O}_i^R > \mathscr{O}_i^{N2} \tag{2}$$

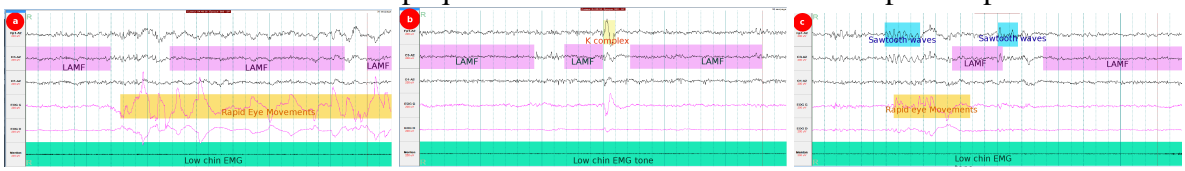
où $\mathscr{O}_i^S, S \in \{W, R, N1, N2, N3\}$ signifie que le stade de sommeil S est observé au cours de l'*époque* i (*époque* courante).
Ces règles sont appliquées dans le système expert avec un moteur à chaînage avant.

3 Résultats

Nous avons testé notre méthode sur les trois *époques* consécutives de la figure 3. Des rectangles colorés indiquent les *grapho-éléments* observés sur ces trois *époques*.

Les observations de la première *époque* (figure 3.a) indiquent clairement un stade R.
Dans la deuxième *époque*, seul le complexe K pourrait indiquer un changement de stade. Comme vu précédemment, d'après la règle G.2 du manuel de l'AASM, l'observation d'un complexe K est un indicateur de début de stade N2. En conséquence, dans la première moitié de cette deuxième *époque* (jusqu'au complexe K), le stade R est observé ; dans la deuxième moitié, le stade N2 est observé (en raison de l'observation du complexe K). Le stade R et le stade N2

FIGURE 3 – Trois époques consécutives toutes scorées R par l'expert



sont donc tous les deux observés à 50% dans cette *époque*, et peuvent donc tous les deux être assignés à cette deuxième *époque*. Pour résoudre ce conflit, la règle I.4 du manuel de l'AASM nous indique que, dans cette situation, le stade R doit être préféré au stade N2. L'*époque* est donc scorée en stade R.
Cette décision est confirmée par les observations de la troisième *époque*, qui comporte clairement des *grapho-éléments* spécifiques du stade R.

4 Discussion

L'exemple traité à la section précédente démontre la pertinence de notre approche, par comparaison aux méthodes utilisées traditionnellement dans la littérature du domaine. Un classifieur tentant de reconnaître le stade de sommeil de la deuxième *époque* de la figure 3 conclurait un stade N2, en raison du complexe K, conformément à ce qu'un expert conclut si on lui présente cette même *époque* de façon isolée.

5 Conclusion

La reconnaissance automatique des stades de sommeil ne doit pas être considéré comme un problème que l'on peut résoudre à l'aide d'un classifieur considérant chaque *époque* de façon isolée. Il est nécessaire de travailler à une échelle de temps plus fine afin de mieux prendre en compte les aspects dynamiques du sommeil. À cette échelle, un système expert semble plus approprié pour prendre une décision. Les règles d'inférence utilisées sont inspirées de celles du manuel de l'AASM. L'observation des experts de l'unité des pathologies du sommeil de la pitié salpêtrière dans leur pratique du scorage des données polysomnographiques a permis une formalisation, fidèle à la pratique, de ces règles. Testée sur une portion de signal comportant une *époque* interprétée avec erreur lorsqu'elle est scorée de façon isolée, notre méthode a permis de prendre la bonne décision.
Dans le futur, nous voulons évaluer la méthode sur des enregistrements polysomnographiques complets puis l'améliorer, en révisant les algorithmes de reconnaissance des *grapho-éléments*, et en intégrant de nouvelles connaissances issues des erreurs qui seront constatées lors de l'évaluation.

Remerciements

Ce travail a été réalisé au sein du Labex SMART avec le support financier de l'état français, représenté par l'ANR, dans le cadre du programme Investissements d'Avenir sous la référence ANR-11-IDEX-0004-02.

Références

ALVAREZ-ESTEVEZ D. & MORET-BONILLO V. (2015). Computer-assisted diagnosis of the sleep apnea-hypopnea syndrome : A review. *Sleep Disorders*, **2015**, 1–33.

ARZT M., YOUNG T., FINN L., SKATRUD J. & BRADLEY T. (2005). Association of sleep-disordered breathing and the occurrence of stroke. *Am J Respir Crit Care Med*, **172**(11), 1447–1451.

BERRY R., BROOKS R., GAMALDO C., HARDING S., LLOYD R., MARCUS C. & VAUGHN B. (2015). *The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology and Technical Specifications, Version 2.2*. American Academy of Sleep Medicine, Darien, Illinois.

DASARATHY B. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, **85**(1), 24–38.

D'AMBROSIO C., BOWMAN T. & MOHSENIN V. (1999). Quality of life in patients with obstructive sleep apnea : Effect of nasal continuous positive airway pressure—a prospective study. *Chest*, **115**(1), 123–129.

ESCOURROU P., MESLIER N., RAFFESTIN B., CLAVEL R., GOMES J., HAZOUARD E., PAQUEREAU J., SIMON I. & ORVOEN FRIJA E. (2010). Quelle approche clinique et quelle procédure diagnostique pour le SAHOS ? *Revue des Maladies Respiratoires*, **27**, S115–S123.

KUNA S., BENCA R., KUSHIDA C., WALSH J., YOUNES M., STALEY B., HANLON A., PACK A., PIEN G. & MALHOTRA A. (2013). Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*, **36**(4), 583–589.

MARIN J., CARRIZO S., VICENTE E. & AGUSTI A. (2005). Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure : an observational study. *Lancet*, **365**(9464), 1046–1053.

PENZEL T., ZHANG X. & FIETZE I. (2013). Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *J Clin Sleep Med*, **9**(1), 89–91.

PUNJABI N. (2008). The epidemiology of adult obstructive sleep apnea. *Proc Am Thorac Soc*, **5**(2), 136–143.

TERÁN-SANTOS J., JIMENEZ-GOMEZ A. & CORDERO-GUEVARA J. (1999). The association between sleep apnea and the risk of traffic accidents. *New England Journal of Medicine*, **340**(11), 847–851.

UEHLI K., MEHTA A., MIEDINGER D., HUG K., SCHINDLER C., HOLSBOER-TRACHSLER E., LEUPPI J. & KÜNZLI N. (2014). Sleep problems and work injuries : A systematic review and meta-analysis. *Sleep Medicine Reviews*, **18**(1), 61–73.

UGON A., GANASCIA J.-G., PHILIPPE C., AMIEL H. & LÉVY P. (2011). How to use symbolic fusion to support the sleep apnea syndrome diagnosis. In M. PELEG, N. LAVRAČ & C. COMBI, Eds., *Artificial Intelligence in Medicine : 13th Conference on Artificial Intelligence in Medicine, AIME 2011, Bled, Slovenia, July 2-6, 2011. Proceedings*, p. 45–54, Berlin, Heidelberg : Springer Berlin Heidelberg.

Approche sémantique pour automatiser le calcul des valeurs nutritionnelles d'une recette de cuisine

Rabia Azzi, Sylvie Despres, Jérôme Nobecourt

UNIVERSITÉ PARIS 13, SORBONNE PARIS CITÉ, LIMICS,(U1142), INSERM, Sorbonne Universités, UPMC Université Paris 6,74 rue Marcel Cachin F-93017 Bobigny cedex, France
prénom.nom@univ-paris13.fr

Résumé : Cet article présente la méthodologie que nous avons développée pour automatiser le calcul des valeurs nutritionnelles (VN) servant à qualifier des recettes de cuisine. Cette méthode de calcul des VN développée par les nutritionnistes, nécessite un appariement entre les ressources textuelles (corpus de recettes) et des données structurées (table de composition nutritionnelle). Nous avons élaboré un méthodologie de calcul composé de quatre étapes. Dans ce papier nous nous focalisons sur les étapes : (1) enrichissement lexical des termes désignant les produits composant les ingrédients ; (2) génération du fichier de calcul à partir d'un patron lexical et interrogation de la table de composition.

Mots-clés : Calcul nutritionel, Corpus, Données structurée, Enrichissement, Patron lexical

1 Introduction

Au cours des 30 dernières années de nombreux travaux scientifiques fondamentaux, cliniques et épidémiologiques ont montré que l'alimentation et les comportements alimentaires constituent des déterminants majeurs pour de nombreuses maladies chroniques non transmissibles (Hercberg, 2013). Cet article présente notre méthodologie pour automatiser le calcul des valeurs nutritionnelles (VN) servant à qualifier des recettes de cuisine. Les ressources utilisées pour ce calcul sont : (1) les tables de composition des aliments ; (2) un corpus brut de recettes de cuisine ; (3) une ressource termino-ontologique (RTO) dans le domaine de la nutrition. Ce travail s'inscrit dans le contexte du développement d'une plateforme, fondée sur l'utilisation de ressources sémantiques pour l'amélioration du suivi nutritionnel de patients atteints de maladies cardiovasculaires. La méthodologie mise en œuvre se décompose en quatre étapes : (1) enrichissement lexical des termes désignant les produits composant les ingrédients ; (2) génération du fichier de calcul à partir d'un patron lexical et de l'interrogation de la table de composition ; (3) calcul et attribution du score ; (4) traduction du score sur une échelle graphique. Dans ce papier nous nous focalisons sur les deux premières étapes. La nature des données manipulées conditionne les deux premières étapes de la méthodologie. L'automatisation du calcul de VN qualifiant une recette, nécessite un appariement précis entre les aliments cités dans les recettes et ceux répertoriés dans les tables de composition. Or un tel appariement est souvent en échec car les vocabulaires utilisés pour décrire ces deux ressources diffèrent. Pour résoudre ce problème, nous proposons d'utiliser une ressource termino-ontologique permettant, outre un meilleur appariement, d'extraire des connaissances supplémentaires utiles au calcul tels que l'état de l'aliment et sa catégorie. En effet, le calcul de la VN (Charrondiere *et al.*, 2011) prend souvent en compte pour chaque ingrédient la manière dont il est préparé (cuit, cru) et la forme sous laquelle il est utilisé (entier, sans peau, etc.).

Cet article est organisé comme suit : la section 2 décrit les ressources utilisées lors des deux

premières étapes de la méthodologie. Les sections 3 et 4 détaillent les deux premières étapes de la méthodologie, ainsi que les résultats expérimentaux obtenus. Enfin, la section 5 conclut cet article et propose des perspectives d'investigations complémentaires.

2 Description des ressources utilisées lors des deux premières étapes de la méthodologie

Techniquement l'approche adoptée est incrémentale. Ainsi, toutes les informations ajoutées restent accessibles à travers les différentes étapes de traitement. La méthodologie opérationnalise des traitements automatiques de la langue (TAL) de base (segmentation, parsing, tokenisation) et de plus haut niveau (extraction d'information). Les principaux traitements sont : (i) segmentation en phrases ; (ii) tokenisation ; (iii) parsing de texte ; (iv) interrogation d'une ressource termino-ontologique ; (v) extraction d'informations ; (vi) analyse morphologique (reconnaissance des aliments dans la base de données). Pour réaliser les étapes 1 et 2 les ressources utilisées par la méthodologie sont : (1) un corpus de recettes ; (2) une table de composition des aliments ; (3) une ressource termino-ontologique.

2.1 Description d'une recette

La recette est un texte semi structuré au format XML. Les éléments pris en compte sont : (1) la liste des ingrédients ; (2) la liste des instructions relatives à la préparation de la recette. A la lecture d'une recette, nous observons que les ingrédients ne figurent pas toujours dans la partie préparation ce qui rend difficile leur reconnaissance. Dans certains cas les ingrédients apparaissent sous forme : (1) verbale dans la préparation (poivrer, au lieu de poivre) ; (2) d'un raccourci (filet de poisson désigné par filet) ; (3) implicite (exemple, l'instruction faire revenir les oignons dans la préparation, déduire que c'est dans l'ingrédient huile d'olive) ; (4) d'un mélange (mélanger les ingrédients de la pâte, etc.). En outre certains ingrédients apparaissant dans la préparation ne figurent pas dans la liste initiale (suggestions d'accompagnement pour le plat). Une première analyse statistique (résultat 0,43%) évaluant le taux d'apparition des ingrédients dans la préparation vient de confirmer les observations *supra*. Une seconde analyse statistique à permis d'estimer à 0,37% le faible taux de reconnaissance des produits composant les ingrédients dans la BD Nutrinet. L'automatisation de ce calcul nécessite un appariement précis entre les produits cités dans les recettes et ceux répertoriés dans les tables de composition. Or, cet appariement est en échec, car les vocabulaires utilisés pour décrire ces deux ressources diffèrent comme le confirme le taux de reconnaissance obtenu.

2.2 Description de la table de composition des aliments

La table de composition des aliments, est constituée d'une liste d'aliments qualifiés par les nutriments les composant. Plusieurs tables de composition des aliments sont accessibles et exploitables sur Internet. Celle qui répond le mieux aux critères de qualités définis par (Greenfield & Southgate, 2007), est la table Nutrinet à partir de laquelle nous avons constitué une base de données (BD) relationnelle.

2.3 Description de la ressource termino-ontologique

Nous avons choisi d'utiliser les modules ALIMENT et PREPARATION de la RTO (exprimée en OWL 2) (Despres, 2014). Le module ALIMENT comporte les concepts (environ 11035) relatifs aux aliments, leurs modes de conditionnement et leurs traits caractéristiques. Ce module permet de répondre aux interrogations concernant les aliments indépendamment de leur utilisation en cuisine. Le module PREPARATION (environ 6164 concepts) décrit principalement les transformations d'aliments *via* les actions culinaires, les types de cuisson et le dressage (par exemple, action de cuisson, découpe, etc.) utilisées dans une recette.

3 Description des deux premières étapes de la méthodologie

Pour exécuter les étapes de la méthodologie d'une manière optimale, nous avons adopté une méthode de traitement à deux couches symboliques : une couche lexicale et une couche de règles. L'objectif est de gérer les spécificités de la recette précédemment identifiés :

- aller au-delà d'une simple recherche lexicale. Par exemple, dans l'expression « cuisse de poulet », seul « poulet » est un « produit », alors que « cuisse » est une « partie de produit » ;
- l'extraction d'ingrédients « implicites » à partir de certains verbes, par exemple, l'action « fariner » correspond à l'ingrédient « farine » ;
- l'expansion des formes. Par exemple, « chantilly [1] » donne « crème chantilly [2] » ;
- l'expansion de termes génériques. Par exemple, le mot « sauce » n'est pas un produit en soi (il peut correspondre à un ingrédient dont les VNs sont renseignées au niveau de la table de composition des aliments ou bien à un aliment composé de plusieurs ingrédients référencé ou non dans la liste d'ingrédients de la recette).

Une étape préliminaire au traitement de la recette qui ne sera pas décrite dans ce papier a permis d'extraire la liste des ingrédients servant à la réalisation du plat. Par exemple, une couche de règle nous a permis de passer du verbe « poivrer » vers l'ingrédient « poivre ». Pour la couche de traitement lexical, nous avons construit un lexique constitué de l'ensemble des labels et altlabels présents dans la RTO.

3.1 Description de l'étape 1

Techniquement, nous avons entamé la première étape de la méthodologie par une segmentation de la recette pour produire une liste d'ingrédients et une liste d'instructions. Chaque ingrédient est représenté selon le patron « produit/quantité/unité ». Par exemple, l'ingrédient « 600g cuisses de poulet » devient « cuisse de poulet/600/g ». Le motif correspondant au « produit » subit un prétraitement linguistique avec TreeTagger [Schmid, 1994]. Cet outil réalise conjointement la tokenisation, la lemmatisation et l'étiquetage morphosyntaxique.

L'enrichissement consiste à exploiter les ressources pour apporter une information supplémentaire aux tokens. Cet enrichissement est réalisé d'une manière plus ou moins profonde selon les tokens considérés. Dans cette partie, l'enrichissement est réalisé avec une recherche lexicale dans la RTO sur la base du lemme de chaque « produit ». Ainsi, nous affectons une étiquette

1. désigne un label dans la RTO
2. désigne un altlabel dans la RTO

lexicale correspondant au concept générique du « produit » dans la RTO. Dans notre exemple (voir figure 1), la recherche du motif « cuisse de poulet » permet d'accéder à la hiérarchie des concepts associée à ce motif, ainsi le motif sera enrichi avec l'étiquette <CABC [3]>cuisse de poulet</CABC>. Comme nous l'avons mentionné précédemment, la profondeur de l'enrichissement dépend du token. L'exploration des relations transversales exprimées sous forme d'« ObjectProperty » dans la RTO (voir figure 1) nous ont permis d'affiner la qualité du motif. Cette RTO décrit des formes plus complexes d'aliments telles que des aliments modifiés (beurre salé), transformés (saumon fumé) ou particularisés (cuisse de poulet). Dans ce travail, les relations utiles sont : « aPourAlimentInitial » et « aPourAlimentOrigine ». Par exemple, si nous suivons la relation « aPourAlimentInitial » associée à <CABC>cuisse de poulet</CABC>, la qualité du motif sera affinée par ajout de nouvelles étiquettes : « cuisse » sera étiquette par « PartieUtile » et « poulet » par « Volaille »(considéré comme produit de base). Dans notre exemple, nous aurons CABC><PUD [4]>cuisse</PUD><VO [5]>poulet</VO></CABC>.

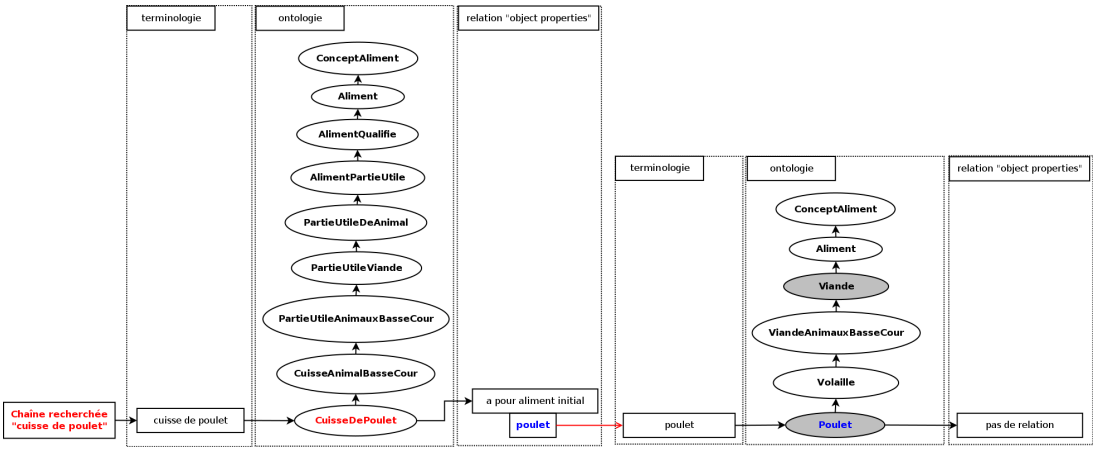


FIGURE 1 – Recherche de cuisse de poulet dans la RTO

Le traitement réalisé sur la préparation débute par le découpage de la préparation en instructions (exemple d'instruction dans la préparation : faites revenir à feu vif les cuisses de poulet). Chaque instruction subit un prétraitement linguistique avec TreeTagger. Des verbes désignant des actions propres au domaine de la nutrition apparaissent dans la partie préparation. Ils ont un impact direct sur les valeurs nutritionnelles de la recette, contrairement aux verbes qui sont en dehors de ce champ sémantique et qui ne doivent pas être pris en compte. L'enrichissement consiste à réaliser un étiquetage de chaque instruction lemmatisée grâce à une recherche lexicale dans la RTO. Le résultat de cette recherche peut être indéterminé et conduire à envisager plusieurs hypothèses. Considérons l'exemple « moule », même si nous nous situons dans un unique contexte (culinaire), une ambiguïté demeure sur le terme « moule ». En effet il peut faire référence à un « ustensile » ou à un « aliment ». Pour résoudre ce problème, nous proposons d'utiliser des modèles estimant la probabilité qu'une étiquette soit affectée à un token et non à

3. CuisseAnimalBasseCour
4. PartieUtileDe
5. Volaille

un autre (par exemple, moule est affecté à Ustensile ou Aliment avec une certaine probabilité en fonction du contexte). Il existe différents modèles adaptés à ce type de problème, que nous ne détaillons pas dans ce papier. Parmi ces derniers figurent les modèles bayésiens et le clustering (Scott *et al.*, 2004). Concernant L'enrichissement des motifs des instructions, nous l'avons réalisé de la même manière que les ingrédients.

Par la suite, pour chaque ingrédient on vérifie la présence du terme qualifiant le « produit de base » dans l'instruction enrichie. Dans le cas où le « produit de base » appartient à une instruction, l'ensemble des termes étiquetés de la même instruction sont associés à ce « produit de base » est par conséquent à l'ingrédient. Dans notre exemple, la recherche du produit de base « poulet » dans l'instruction « <ADF [6]>faire</ADF><ADC [7]>revenir</ADC> <FDC [8]>feu vif</FDC><CABC><PUD>cuisse</PUD><VO>poulet</VO></CABC> donne un résultat positif. Ainsi le produit « cuisse de poulet » reçoit l'ensemble des items de l'instruction. Afin d'identifier l'état du produit dans la recette et la catégorie alimentaire du produit qui sont nécessaires pour le calcul des VNs. Nous avons procédé en deux étapes : (1) pour l'état de produit, vérification de la présence d'une action de cuisson dans la liste des items de l'instruction enrichie associée a ce produit. Si la réponse est positive alors l'état de l'aliment est « cuit », dans le cas contraire il est « cru ». Dans notre exemple, « revenir » correspond à une action de cuisson puisqu'il correspond à un concept spécifique du concept générique « ActionDeCuisson » (revenir est classé sous action de cuisson) ; (2) pour la catégorie alimentaire du produit, nous nous appuyons sur la RTO pour identifier les concepts génériques prédéfinis pour les catégories d'aliment (« Fruit », « Légume », « Viande », etc.). L'interrogation est de type étiquette du « produit de base » est-elle sous concept générique de « Fruit », « Légume », « Viande », etc. Tant que la réponse n'est pas positive, l'interrogation se poursuit jusqu'à l'identification de la catégorie. Dans notre exemple, la catégorie pour « poulet » est « viande » puisque l'étiquette <Volaille> a pour concept générique <Viande>.

3.2 Description l'étape 2

La deuxième étape de la méthodologie consiste à produire un fichier de calcul structuré sous forme d'un patron « identifiant produit/catégorie/poids ». Cet identifiant produit correspond à celui figurant dans la BD Nutrinet. En partant de la liste des produits enrichis dans la première étape, une chaîne de caractère est construite selon le patron « produit de base/partie du produit/action de cuisson/autre modification ». En d'autres termes : (1) si dans la liste des termes étiquetés associés au produit apparaît un des motifs du patron, le motif du patron prendra la valeur du terme étiqueté ; (2) si après avoir testé tous les termes du produit enrichi aucun résultat n'a été identifié, le «motif du patron» prendra la valeur par défaut (par exemple, pour la partie du produit la valeur par défaut est « entier »). Dans notre exemple, pour le produit enrichi « cuisse de poulet » le patron sera ainsi construit : (1) le produit de base -> poulet ; (2) la partie du produit -> cuisse puisque son étiquette est « PartieUtileDe » ; (3) action de cuisson -> revenir puisque sont étiquette est « ActionDeCuisson » ; autre modification -> feu vif puisque l'étiquette est « ForceDeCuisson ». Finalement le patron pour le produit « cuisse de poulet »

6. ActionDeFabrication
7. ActionDeCuisson
8. ForceDeCuisson

sera « poulet/cuisse/revenir/feu vif ».

Une requête SQL sur la base de données est ensuite formulée, pour récupérer toutes les lignes dans lesquelles apparaît la chaîne « produit de base ». Le résultat de cette requête constitue la liste de candidats potentiels pour le produit. Ensuite cette liste est réduite jusqu'à l'identification du produit. La solution consiste à utiliser un comparateur pour spécifier l'ordre du tri. Chaque fois que la méthode de tri doit comparer deux éléments, elle utilise un token du patron précédemment construit (produit de base/partie du produit/action de cuisson/autre modification). Par exemple, le premier comparateur sera « produit de base », le second « partie du produit », etc.

Les unités de chaque produit sont converties en gramme, grâce à une table de conversion fournie par l'EREN, pour obtenir les nouvelles quantités. La procédure est la suivante : (1) si l'unité du « produit de base » est déjà en gramme alors la valeur du poids ne change pas ; (2) si l'unité du « produit de base » n'est pas exprimée en gramme, une table de conversion de calibre du produit intervient. Par exemple, l'ingrédient « poivron/3/unité ». Le poids moyen d'un poivron est de 50g, le poids en gramme du poivron dans la recette vaut 3x50=150g. A l'issue de ce traitement, chaque produit sera qualifié avec un poids en gramme. Enfin le fichier de calcul est établi et est utilisé pour le calcul du score.

4 Résultats et discussion

D'une manière générale, même avec une précision moindre, on obtient la distinction entre « cru » et « cuit » qui est essentielle pour obtenir des valeurs correctes en entrée du calcul. En ce qui concerne la qualité de l'appariement entre les produits de la recette et leurs entrées dans la table de composition, les résultats obtenus sont prometteurs et s'avèrent même concluants dans certains cas. Cet appariement n'est pour l'instant pas quantifié. Il est clair que l'ordre exact des entités lexicales du patron « produit de base/partie du produit/action de cuisson/autre modification » construit à l'étape 2 est très important. En effet chaque entité lexicale a un impact direct sur les valeurs nutritionnelles du produit dans le plat final (plat préparé) conduisant à des questions du type « est ce que la partie de l'aliment est plus importante que l'action de cuisson, etc ». Pour répondre à ces interrogations, il est primordial d'enrichir les connaissances dans le domaine de la nutrition pour gagner en précision dans l'appariement. Une fois la méthodologie modifiée dans ce sens nous avons prévu de réaliser une évaluation plus représentative sur un corpus de recettes.

5 Conclusion

Nous avons présenté dans cet article notre approche pour réaliser les étapes 1 et 2 de la méthodologie de calcul automatique des valeurs nutritionnelles qualifiant des recettes de cuisine. Ce travail souligne l'apport des ressources sémantiques termino-ontologiques pour la résolution des problèmes nécessitant le recours à des connaissances implicites. En effet, ces ressources permettent une meilleure précision lors des traitements purement algorithmiques ultérieurs. Il est clair qu'il nous reste à approfondir les traitements pour tirer parti de la séquentialité des unités locales des énoncés, en utilisant des modèles adaptés comme les Modèles de Markov Cachés (Lafferty *et al.*, 2001).

Références

AZZI R. (2014). Conception d'une base de données à partir de la table de composition des aliments nutrinet. In *Rapport de stage M1 Informatique Biomédicale. Université Paris 13, Sorbonne Paris Cité, LIMICS, INSERM,(UMRS 1142)*.

CHARRONDIERE U.-R., BURLINGAME B., BERMAN S. & ELMADFA I. (2011). Food composition study guide-answers to questions and exercises.vol2. In *the international network of food data system,FAO, Rome, Italy, 2011*.

DESPRES S. (2014). Construction d'une ontologie modulaire pour l'univers de la cuisine numérique. In *25èmes Journées francophones d'Ingénierie des Connaissances, May 2014, Clermont-Ferrand, France*, p. 27–38.

GREENFIELD H. & SOUTHGATE D.-A.-T. (2007). Données sur la composition des aliments : production, gestion et utilisation. In *Organisation des Nations Unies pour l'alimentation et l'agriculture,FAO, Rome, Italie*.

HERCBERG S. (2013). Propositions pour un nouvel élan de la politique nutritionnelle française de santé publique. In *Rapport 15/11/2013 pour Mme la ministre de la santé, dans le cadre de la Stratégie Nationale de Santé, France*.

LAFFERTY J., MCCALLUM A. & C.N.PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML,Williamstown, MA, USA.2001*, p. 282–289.

NUTRINET-SANTÉ E. (2013). table de composition des aliments. In *France :Economica*, p. 296.

SCOTT M., GUINNESS J. & ZAMANIAN A. (2004). Name tagging with word clusters and discriminative training. In *Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT/NAACL-04), Boston, USA*, p. 337–342.

OMIConto : une ressource termino-ontologique pour la qualification et l'indexation des outils d'analyse en sciences *omiques*

Vincent J. Henry[1], Lina F. Soualmia[2,3], Julien Grosjean[2], Arnaud Desfeux[4], Stéfan J. Darmoni[2,3], Bruno J. Gonzalez[5]

[1] OMICtools project
vincent643@gmail.com
[2] NormaSTIC CNRS 3638, LITIS EA 4108, Normandie Universités, Univ. Rouen
{Lina.Soualmia, Julien.Grosjean, Stefan.Darmoni}@chu-rouen.fr
http://www.hetop.eu/hetop/
[3]LIMICS, INSERM UMR_1142, Sorbonne Universités, Paris
[4]omicX, Sotteville-lès-Rouen
arnaud.desfeux@omictools.com
http://omictools.com/
[5] INSERM ERI-28, IRIB, Normandie Universités, Univ. Rouen
bruno.gonzalez@univ-rouen.fr

Résumé : Les progrès récents dans les technologies *omiques* ont créé des opportunités sans précédent pour la recherche biomédicale. Le défi actuel réside dans l'analyse et l'interprétation des données produites grâce aux outils bioinformatiques (logiciels et bases de données). La base de données OMICtools a été créée dans le but d'indexer et classifier ces outils accessibles de manière diffuse. Lors de son développement, l'absence de vocabulaire contrôlé pour définir et rechercher les outils a été soulignée. Afin de répondre à ce besoin, nous avons développé la ressource termino-ontologique (RTO) OMIConto. Ainsi, 909 termes hiérarchisés ont été intégrés dans le portail HeTOP (Health Terminologies and Ontologies Portal). Depuis, l'enrichissement et la formalisation d'OMIConto a permis le développement de 1 366 concepts en relation appartenant à 3 concepts de haut niveau : "Structure", "Index" et "Metainformation". Les alignements avec d'autres RTO présentent un taux de 45% de termes spécifiques à OMIConto. Les premiers résultats de l'utilisation conjointe d'OMIConto et de HeTOP suggèrent une recherche efficace des ressources associées au domaine des sciences *omiques* dans la base de données bibliographique MEDLINE. À terme, la formalisation du vocabulaire contrôlé devrait permettre d'implémenter un moteur de recherche performant interne à la base de données OMICtools.

Mots-clés : Terminologie, Ontologie, Sciences omiques, MeSH, Indexation, Recherche d'information.

1 Introduction

Depuis une dizaine d'années, les progrès permanents en informatique et dans les technologies de production de données biologiques à haut débit (séquençage, puces, spectrométrie de masse, etc.) pour les sciences *omiques* (génomique, transcriptomique, protéomique, métabolomique, etc.) ont révolutionné les sciences de la vie (Hoheisel, 2006 ; Metzker, 2010 ; Nilsson et al., 2010). Ces sauts technologiques ont entraîné un changement d'échelle aux niveaux de la production, de l'analyse et du stockage des données, à l'origine de la création de structures et plateformes dédiées. L'avènement de prochains progrès technologiques devrait être à l'origine d'une nouvelle vague de démocratisation des approches *omiques*, libérées des contraintes de stockage et de coût (Kilianski et al., 2012). Ainsi, la

colonisation prédite des laboratoires de recherche et des hôpitaux confrontera des chercheurs et cliniciens issus d'autres champs d'expertise aux sciences *omiques* (les néo-utilisateurs).

Le défi actuel réside dans la capacité de la communauté scientifique biomédicale à répondre d'une façon satisfaisante à l'analyse et à l'interprétation de l'avalanche des données massives générées (Nekrutenko & Taylor, 2012). Cette exploitation repose sur un large éventail d'outils bioinformatiques, notamment des logiciels et des bases de données (BdD). De fait, chaque jour, de nouveaux outils sont mis à la disposition de la communauté des sciences de la vie et présentés par exemple dans des articles scientifiques (plus de 2 000 par an depuis 2012). Cependant, ces outils sont mal indexés, peu documentés et dispersés (Dellavalle et al., 2003). L'absence de méthodes d'indexation rend difficile la recherche d'information associée à ces ressources et limite ainsi leur diffusion au sein de la communauté bioinformatique confirmée et par extension à la communauté des néo-utilisateurs. Dans ce contexte, plusieurs appels à organiser ces ressources ont été lancés (Cannata et al, 2005).

Des approches axées sur la recherche d'information ont été développées afin de trouver directement les ressources existantes dans le Medical Literature Analysis and Retrieval System Online (MEDLINE). Les documents sont indexés par le thésaurus Medical Subject Headings (MeSH) (Nelson et al. 2001) et leur recherche est facilitée par le moteur intégré Public MEDLINE (PubMed). Si les termes MeSH sont peu adaptés pour indexer les articles d'outils bioinformatiques, des combinaisons d'analyse de texte et d'utilisation de PubMed et du MeSH ont montré des preuves de concepts intéressantes pour palier ce déficit (Yamamoto & Takagi, 2007 ; De la Calle et al., 2009 ; Duck et al., 2013). Malheureusement, ces étapes préliminaires n'ont pas été confirmées à plus grande échelle.

D'autres approches rassemblent les outils dans de nouvelles BdD. Il est possible de distinguer deux sortes d'annuaire : a) des BdD spécifiques qui recensent les outils dans un domaine précis de façon presque exhaustive ; Réseaux de signalisation protéiques (Pathguide), métabolites (MetaChemBio) ou ARN non codant (ncRNA DB) (Klingström T, Plewczynski, 2011 ; Paschoal et al., 2012 ; Minkiewicz et al., 2015) ; des BdD à plus larges spectres de domaines, qui misent sur un travail collaboratif de renseignement (Artimo et al., 2012 ; Brazas et al., 2012 ; Li et al. 2012 ; Henry et al., 2014 ; Ison et al., 2016). L'ensemble de ces efforts d'organisation présente des limites liées, respectivement, à leur hyper-spécialisation ou leur difficulté à être à jour.

Un de ces annuaires (Bio.tool ; Ison et al., 2016) s'appuie sur une ontologie (EMBRACE Data and Methods : EDAM) qui a pour domaine la gestion des données biologiques générées informatiquement (Ison et al., 2013). Cette ontologie de 3 220 concepts se focalise sur les types de données et l'aspect technique de leur manipulation (langage de programmation) ou de leur stockage (formats de fichier) excluant les technologies à l'origine de leur production. Ainsi, le couple bioregistry-EDAM est un outil axé sur l'aspect informatique et se destine aux utilisateurs confirmés.

OMICtools (accessible gratuitement à l'URL http://omictools.com) propose une classification destinée aux néo-utilisateurs. En effet, le parti pris du projet est d'utiliser une classification des outils à partir des technologies de production de données (séquençage à haut-débit, puces, FIG. 1Aa1), des types d'études biologiques (génomique, transcriptomique, FIG. 1Aa2) ou encore des domaines d'intérêt (néoplasme, *drug discovery*). La partie technologie est subdivisée en application par technologie puis étapes d'analyses (respectivement : *whole exome sequencing*, *Copy Number Variation (CNV) detection*, FIG. 1Ab et FIG. 1Ac). La partie d'études biologiques est subdivisée en sous types d'analyse, puis logiciel ou BdD d'intérêt (respectivement : étude de variation génomique et BdD associant pathologie et variation génétique). Ainsi, plus de 10 000 outils sont classés en 909 catégories hiérarchisées. Enfin, les outils sont également rattachés à des métadonnées fonctionnelles renseignant sur leur nature et les aspects techniques (système d'exploitation, interface, format). Cependant, ces catégories et ces métadonnées ne sont pas reliées entre elles et manquent de formalisme, d'homogénéité et de structuration. De fait, la classification d'OMICtools ne constitue pas un vocabulaire contrôlé et ne possède pas les caractéristiques nécessaires à une indexation utile pour une recherche d'information efficace.

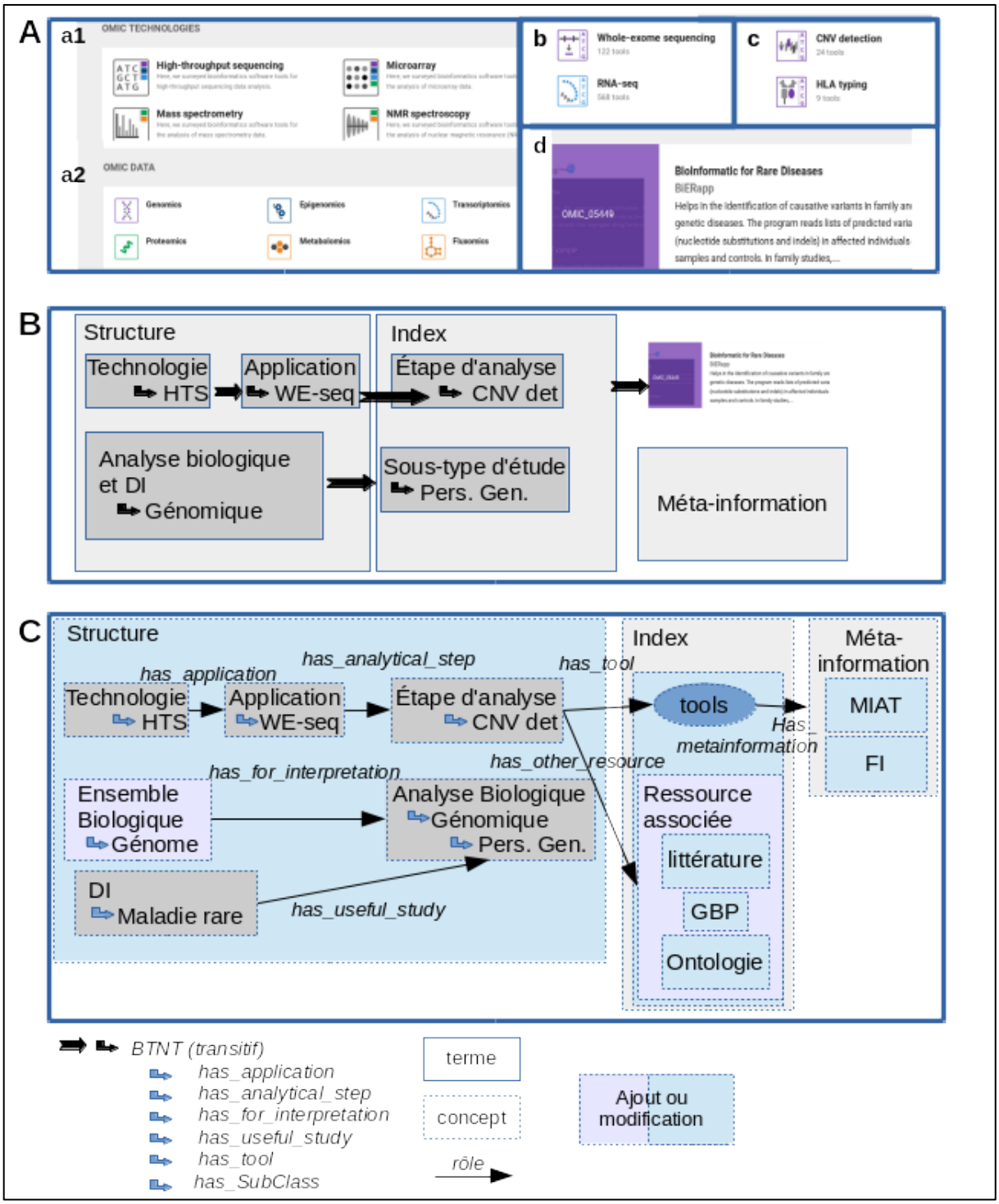


FIGURE 1 – Évolution et modélisation d'OMIConto. (A) Illustration de la classification d'OMICtools : a1) par technologie, a2) par type d'étude ou domaine, b) par application, c) par étape d'analyse et d) outil. (B) Extraction terminologique de l'architecture d'OMICtools, les termes de l'index permettent d'indexer les outils. (C) Représentation du modèle ontologique. Les outils sont des instances de « tools ». DI : domaine d'intérêt ; GBP : Guide de bonnes pratiques, MIAT : minimimum information about tools ; CNV det : copy number variation detection ; Pars. Gen. : personal genomics ; HTS : Séquençage à haut débit ; WE-seq : whole exome sequencing.

L'objectif prochain d'OMICtools est d'intégrer un moteur de recherche efficace pour les visiteurs de la BdD mais aussi de proposer une solution de recherche externe pour son alimentation par les curateurs. Pour mener à bien ce projet, il est essentiel que le système d'information puisse s'appuyer sur une RTO solide alliant la richesse et le contrôle sémantique de la terminologie et le formalisme de l'ontologie.

Dans cet article, nous expliquons comment nous avons extrait les termes originaux d'OMIConto à partir de la classification d'OMICtools (améliorée au fil de l'investissement de la communauté de curateurs et du retour des utilisateurs), puis étendu ces termes en concepts reliés, permettant l'indexation de la BdD d'OMICtools. Enfin, nous verrons l'intérêt de son intégration dans le portail multi-terminologie de santé HeTOP Grosjean et al., 2013), notamment pour enrichir l'interrogation de MEDLINE.

2 Approche proposée

Le développement de la partie terminologique d'OMIConto (FIG. 1B) s'appuie sur le guide des bonnes pratiques de Tao et collaborateurs (Tao et al., 2013). La première étape consiste à définir les termes suite à l'extraction des catégories de la classification des outils dans OMICtools. La définition de ces termes implique d'y associer un libellé préféré, des synonymes, des acronymes et une définition. La deuxième étape consiste à attribuer des relations hiérarchiques de termes plus larges ou plus précis (*broader term* ou *narrower term* ; BTNT) reflétant ceux définis entre les catégories d'OMICtools. L'ensemble de ces étapes a été facilité grâce à l'intégration des termes initiaux dans HeTOP (FIG. 2). De plus, cette intégration a permis de déterminer des liens externes vers des termes équivalents provenant d'autres RTO, ceci grâce à la production d'alignements automatiques par un algorithme combinant approche syntaxique et sémantique (Mérabti et al., 2012) suivie par une validation manuelle : deux RTO biomédicales classées parmi les plus consultées dans BioPortal (Noy et al., 2009), le National Cancer Institute thesaurus (NCIt) (Sioutous et al., 2007) et l'Ontology for Biomedical Investigation (OBI) (Dugan et al., 2014) (dont une partie de leur domaine s'intéresse à la médecine de précision), la Gene Ontology (GO) qui compile les fonctions biologiques utilisées pour l'annotation des gènes (The Gene Ontology Consortium, 2000), le MeSH (Nelson et al., 2001) et EDAM ontology (Ison et al., 2013).

L'étape de formalisation en OWL2 (FIG. 1C) a consisté à extraire les concepts à partir des termes précédents et des rôles à partir de leurs relations BTNT disponibles dans HeTOP vers l'éditeur d'ontologies Protégé 5.0 (Kapoor & Sharma, 2010). Dans un premier temps, une super-classe « `outil` » contenant les types d'outil d'analyse est créée. Puis, les relations BTNT sont conservées sous forme transitive et des sous-propriétés non-transitives sont définies. Ainsi les relations `has_application`, `has_analytical_step` ou `has_tool` permettent une interrogation plus fine de l'ontologie. À l'inverse, de nouvelles relations `has_useful_study` sont créées et deviennent des sous-classes de BTNT. Enfin des relations `has_metainformation` et leur sous-ensemble (`has_fomat`, `has_language`) associent les outils à leur aspect technique. Ces différentes étapes ont permis de définir un modèle définitif cumulant des caractéristiques terminologiques et ontologiques.

3 Résultats

3.1 Modèle et description

L'extraction des termes de la classification de OMICtools (FIG. 1A) et la modélisation de la terminologie dérivée ont permis de définir 3 types de concepts : structure, index et méta-information (FIG. 1B). La plupart des concepts terminologiques (80%) possèdent une définition (FIG. 2). Le pourcentage d'alignement des concepts avec d'autres RTO est de 56,25 %. Cet alignement permet d'augmenter la couverture de recherche pour les termes

concernés. Globalement, nous obtenons une répartition assez homogène (NCIt & OBI : 24 %, GO : 20 %, MeSH 27 % et EDAM 29%). Par contre, chacune de ces RTO présente un alignement particulier avec OMIConto. Les « `étapes d'analyses` » sont alignées à 75 % avec EDAM, les « `technologies et applications` » sont alignées à 56 % avec NCIt & OBI et 29 % avec le MeSH. Enfin, la GO s'aligne exclusivement avec des concepts de « `systèmes biologiques` » pour qui l'alignement se partage principalement entre la GO (34 %) et le MeSH (33%).

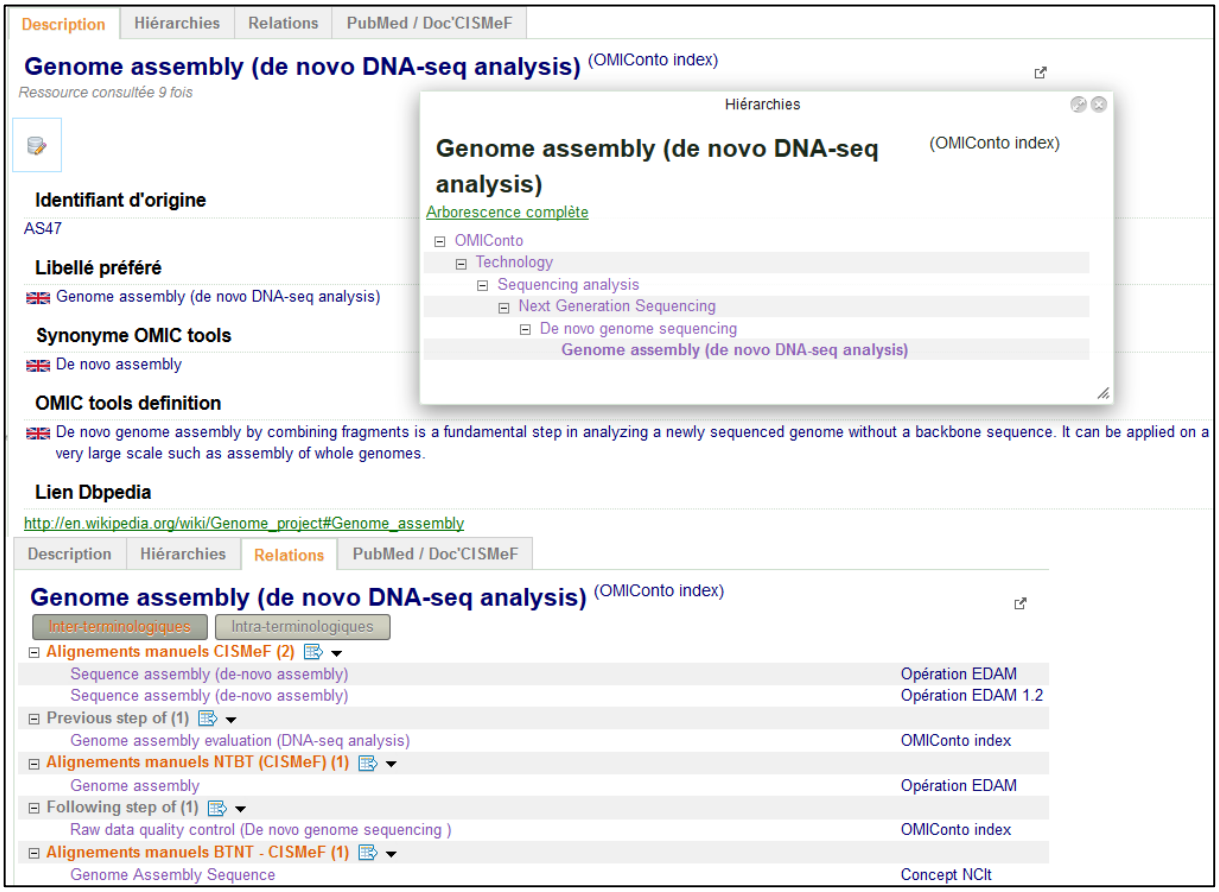


FIGURE 2 – HeTOP. Illustration synthétique des réponses données par HeTOP suite à la recherche des termes correspondant à '`genome assembly`'. Les réponses sont classées en *description*, *hiérarchies* et *relations*.

La formalisation ontologique à nécessité l'ajout de 400 concepts impliquant la création d'une super-classe « `ressource` » (remplaçant `index` pour l'instanciation des outils), deux sous-classes de « `structures` » (« `ensemble biologique` » et « `domaine d'intérêt` ») au sein desquelles vient s'ajouter l'index de la terminologie (FIG. 2C). Au total, OMIConto représente 1 366 concepts dont 909 sont adaptés à l'indexation d'outils. Au final, les concepts de « `technologies` » (séquençage à haut débit) sont reliés en aval à des « `applications` » (*whole exome sequencing*) qui sont reliées à des « `étapes d'analyses` » susceptibles de transformer des données brutes en données interprétables (*CNV detection*, FIG. 1C). Ces concepts ont le même type de relation vers des étapes « `d'analyse biologique` » (génomique). Les concepts « `d'analyses biologiques` » sont reliés en amont aux ensembles biologiques (génome) et aux domaines d'intérêts (maladie rare). L'ensemble de ces concepts représente l'axe «`rôle`» des outils. L'axe « `aspect technique` » est assuré par les concepts de « `Métainformation` » (langage, input data, auteur).

3.2 Utilisation dans HeTOP

Pour chaque terme, l'intégration d'OMIConto dans HeTOP, permet l'accès rapide et la visualisation de sa hiérarchie, de ses synonymes, acronymes, de sa définition et de ses relations internes ou avec d'autres RTO contenues dans la BdD (FIG. 2). Une autre option d'HeTOP est la génération automatique de requêtes dans la syntaxe de PubMed à partir du terme d'intérêt et de ses synonymes, acronymes et alignements. Les résultats de cette requête peuvent être améliorés grâce à un filtre qui prend en compte les caractéristiques d'indexation du MeSH. Ainsi, des résultats préliminaires montrent qu'il est possible d'augmenter la précision lexicale de la recherche d'information sans perdre en rappel (moins de réponses avec une proportion équivalente correspondante au sujet d'intérêt comme illustré dans la FIG. 3).

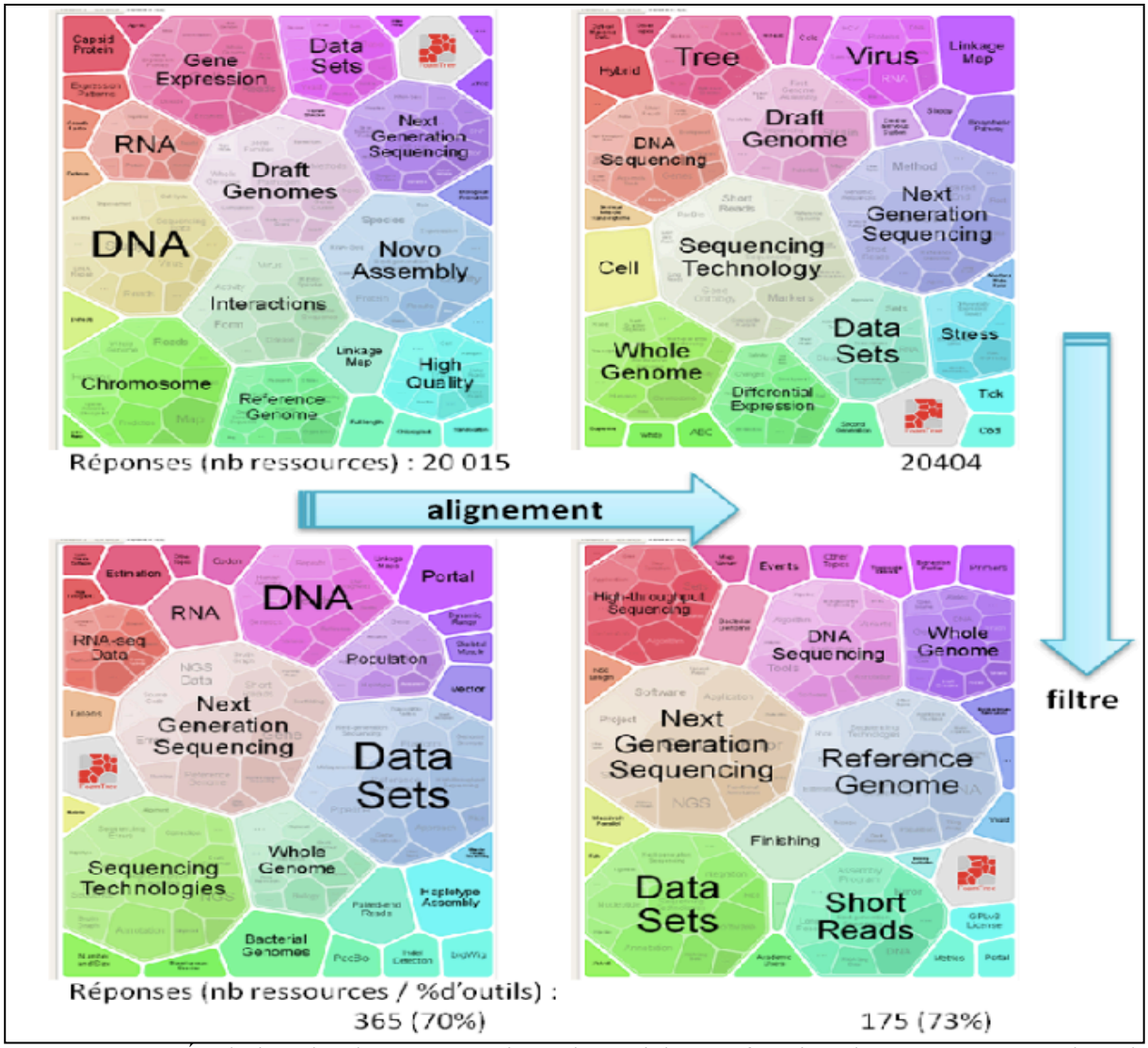


FIGURE 3 – Évolution du champ sémantique des articles en fonction des requêtes sur PubMed. Représentation des mots ou groupes de mots statistiquement représentatifs à partir d'une requête (`genome assembly`) en fonction de l'emploi d'alignements et/ou de l'emploi de filtres dans les requêtes. L'emploi de l'alignement réduit le nombre de réponses à la requête. L'utilisation du filtre enrichit le nombre de réponses correspondant à des outils.

4 Discussion

La formalisation de la classification établie pour le site OMICtools a permis la création de la RTO OMIConto permettant d'indexer sa BdD. Elle comprend 624 concepts dans la partie « `tools` » adaptés à l'indexation des outils et reliés à leur technologie, application, type d'étude biologiques ou encore domaine d'intérêt. OMIConto est donc susceptible de servir de base à l'implémentation d'un moteur de recherche interne à la BdD OMICtools. De plus l'extension ontologique de la classification initiale permet d'envisager l'indexation de nouvelles ressources comme des articles de revue, de méthodes, des guides de bonnes pratiques ou d'autres ontologies qui pourront être traitées comme les outils (FIG. 1).

Lors de l'extraction terminologique, la recherche de définitions pertinentes s'est heurtée aux mêmes difficultés que la curation des ressources de la BdD. En effet, l'absence de vocabulaire contrôlé dans le domaine des sciences *omiques* se répercute autant sur les outils que sur la définition des concepts. De même, si la proportion d'alignements vers d'autres RTO est homogène, une exploration par concept de haut-niveau permet de distinguer des spécificités et montre une répartition des centres d'intérêts entre les thématiques. Finalement, OMIConto présente ses propres termes et centralise diverses approches pour une nouvelle proposition de modèle. Ce résultat suggère que la méthodologie de centralisation des ressources bioinformatiques utilisée par OMICtools pourrait être appliquée à la centralisation des RTO susceptibles de les indexer.

En parallèle, la conception d'OMIConto à partir d'une classification préexistante rend consistants les efforts précédents de classification. L'adaptation en terminologie puis la formalisation ontologique a nécessité l'utilisation de modèles successifs afin de se détacher suffisamment de la classification initiale sans toutefois la dénaturer. La réflexion sur la pertinence des concepts, leur consistance et leur homogénéité a donc été à l'origine de modifications structurelles entre la classification, la terminologie puis l'ontologie. Ainsi, de nouvelles classes de haut-niveau ont du être conçues pour la consistance et le formalisme.

Enfin, les premiers essais de requêtes vers MEDLINE à travers HeTOP *via* PubMed sont encourageants. Si ces résultats nécessitent une évaluation formelle, ils montrent la possibilité d'une première étape vers l'obtention d'une réponse pertinente et démontrent l'intérêt de combiner une approche algorithmique multi-terminologique et une approche de filtration des requêtes à partir de la maîtrise du domaine et de l'indexation par le MeSH. Dans ce contexte, OMIConto pourrait donc s'avérer très utile comme aide à la curation de la BdD.

Références

ARTIMO P. et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*, 40(Web Server issue):W597–603.

BRAZAS MD. et al. (2012). A decade of Web Server updates at the Bioinformatics Links Directory: 2003-2012. *Nucleic Acids Res* , 40(Web Server issue):W3–W12.

CANNATA N. et al. (2005). Time to organize the bioinformatics resourceome. *PLoS Comput Biol*, 1:e76.

DE LA CALLE G. et al. (2009). BIRI: a new approach for automatically discovering and indexing available public bioinformatics resources from the literature. *BMC Bioinformatics*, 10:320.

DELLAVALLE RP. et al. (2003). Information science. Going, going, gone: lost Internet references. *Science*, 302:787–788.

DUGAN VG. et al. (2014). Standardized metadata for human pathogen/vector genomic sequences. *PloS One* , 9:e99979.

DUCK G. et al. (2013). bioNerDS: exploring bioinformatics' database and software use through literature mining. *BMC Bioinformatics*, 14:194.

GROSJEAN J. et al. (2011). Health multi-terminology portal: a semantic added-value for patient safety. *Stud Health Technol Inform*, 166:129–138.

HENRY VJ. et al. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database J Biol Databases Curation*.

HOHEISEL JD. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, 7:200–210.

ISON J. et al. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinforma Oxf Engl*, 29:1325–1332.

ISON J. et al. (2016). Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. Jan 4; 44(D1):D38-47.

KAPOOR, B., & SHARMA, S. (2010). A comparative study ontology building tools for semantic web applications. *International journal of Web & Semantic Technology* (IJWesT).

KILIANSKI A. et al. (2015). Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience*, 4:12.

KLINGSTRÖM T. & PLEWCZYNSKI D. (2011). Protein-protein interaction and pathway databases, a graphical review. *Brief Bioinform*, 12:702–713.

LI J-W. et al. (2012). The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. *Nucleic Acids Res*, 40(Database issue):D1313–1317.

MÉRABTI T. et al. (2012). Aligning biomedical terminologies in French: towards semantic interoperability in medical applications. In book: Medical Informatics, 41–68, InTech Press.

METZKER ML. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11:31–46.

NEKRUTENKO A. &, Taylor J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet*, 13:667–672.

NELSON SJ. et al. (2001). Relationships in Medical Subject Heading. Relationships in the Organization of Knowledge, eds. Kluwer Academic Publishers, p. 171–184.

NILSSON T. et al. (2010). Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods*, 7:681–685.

NOY NF. et al. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*. Jul;37(Web Server issue):W170-3.

MINKIEWICZ . P. et al. (2015). Using Internet Databases for Food Science Organic Chemistry Students To Discover Chemical Compound Information. *J Chem Educ*, 92:874–876.

PASCHOAL AR. et al. (2012). Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases. *RNA Biol* , 9:274–282.

TAO C. et al. (2013). Terminology representation guidelines for biomedical ontologies in the semantic web notations. *J Biomed Inform*, 46:128–138.

YAMAMOTO Y. & TAKAGI T. (2007). OReFiL: an online resource finder for life sciences. *BMC Bioinformatics*, 8:287.

MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein

Solène Eholié[1], Mike Donald Tapi Nzali[1,2],
Sandra Bringay[1], Clement Jonquet[1,3]

[1] LABORATOIRE D'INFORMATIQUE, DE ROBOTIQUE ET DE MICROÉLECTRONIQUE DE MONTPELLIER (LIRMM), Université de Montpellier, France
prenom.nom@lirmm.fr

[2] INSTITUT MONTPELLIÉRAIN ALEXANDER GROTHENDIECK (IMAG), Université de Montpellier, France
mike-donald.tapi-nzali@univ-montp2.fr

[3] CENTER FOR BIOMEDICAL INFORMATICS RESEARCH (BMIR), Stanford University, USA

Abstract : Il existe un écart notable à la fois d'ordre lexical et sémantique entre le vocabulaire des professionnels de la santé et celui des patients. À notre connaissance, il n'existe pas de ressource formalisée pour le français liant ces deux niveaux de vocabulaire. Nous présentons dans ce travail, une formalisation en SKOS d'un vocabulaire reliant ces deux niveaux d'expertise dans le cadre de la thématique du cancer du sein ainsi qu'une méthode d'alignement de la terminologie résultante, MuEVo, à des terminologies biomédicales de référence à savoir MeSH, SNOMED et MedDRA.

Mots-clés : Système d'organisation des connaissances (SOC), terminologies biomédicales, vocabulaire patient

1 Introduction et motivations

Selon une étude de TNS Sofres[1] réalisée en 2013, un Français sur deux a déjà recherché ou échangé des informations sur sa santé via le Web. De même, on trouve en ligne de très nombreuses publications scientifiques produites par les professionnels de santé comme dans la base bibliographique PubMed[2]. Des données en très grande quantité sont donc disponibles sur des sujets médicaux décrits selon deux niveaux d'expertise : patient et médecin.

Il existe de nombreuses formalisations du vocabulaire médecin en français, sous la forme de terminologies comme MeSH, SNOMED ou MedDRA[3]. Cependant, McCray *et al.* (1999) ont montré qu'il existe un écart notable entre le vocabulaire scientifique et technique utilisé par les médecins et celui vulgarisé des patients. Cet écart lexical et/ou sémantique handicape par exemple les patients dans leur recherche d'informations médicales (Kogan *et al.*, 2001). Certains travaux (Zeng & Tse, 2006; Jiang *et al.*, 2013) se sont donc intéressés à la création de CHV[4] (Consumer Health Vocabulary).

Le travail présenté dans cet article fait suite aux travaux de Tapi Nzali *et al.* (2015) qui

[1] http://www.patientsandweb.com/wp-content/uploads/2013/04/A-la-recherche-du-ePatient-externe.pdf

[2] http://www.ncbi.nlm.nih.gov/pubmed

[3] http://mesh.inserm.fr/mesh/, http://www.meddra.org

[4] vocabulaire composé d'un ensemble de termes utilisés par les non-experts (patients, leurs familles, etc.) pour exprimer des concepts médicaux

ont proposé une méthode originale de construction *semi-automatique* d'un vocabulaire patient/médecin à partir des médias sociaux. Ce vocabulaire que nous formalisons dans cet article et appelons MuEVo, est spécifique à la thématique du cancer du sein. Notre objectif est maintenant de construire une ressource structurée, exploitable par une machine et conforme aux standards du Web sémantique, qui permettra de faire le pont entre ces niveaux d'expertise, afin de pouvoir effectuer des traitements automatiques (e.g., recherche d'information (Zarro & Lin, 2011), classification automatique).

Nous proposons dans cet article une formalisation en SKOS[5] du vocabulaire MuEVo (section 2) puis une méthodologie pour l'aligner avec les terminologies présentes dans le serveur de terminologies francophones, SIFR BioPortal (section 3) (Jonquet *et al.*, 2016).

2 Formalisation SKOS

2.1 Présentation des données

Pour valider ce travail préliminaire, nous avons utilisé 173 relations entre termes patient/médecin, spécialisées sur le cancer du sein, issues de (Tapi Nzali *et al.*, 2015). Ces relations ont été obtenues via un alignement entre un corpus patient constitué d'un ensemble de messages extraits de médias sociaux (forums[6] et groupes Facebook publics[7]) et un vocabulaire médecin cible à savoir la liste de termes de référence proposée par l'INCa[8] (Delavigne, 2012). À chaque relation, sont associés un type (*abréviation*, *erreur d'orthographe* ou *association*), la méthode utilisée pour détecter la relation et un poids assigné par la méthode. Le tableau 1 présente quelques exemples de relations.

Terme patient	Terme médecin	Type de relation
nez	pharynx	association
abaltion	ablation	erreur d'orthographe
onco	oncologue	abréviation
traitement hormonal	hormonothérapie	association

Tableau 1 : Exemples de relations patient/médecin extraits de Tapi Nzali *et al.* (2015)

2.2 Spécification du modèle

SKOS est une recommandation du W3C pour représenter des vocabulaires controlés (Miles *et al.*, 2005). C'est un standard très utilisé dans la communauté du Web sémantique. Le thésaurus AGROVOC[9] par exemple est formalisé en SKOS. L'unité de connaissance en SKOS est le *skos:Concept*. Un *skos:Concept* est une ressource RDF qui formalise une idée, une réalité.

[5] Simple Knowledge Organization System - https://www.w3.org/2004/02/skos/

[6] *cancerdusein.org*

[7] *Cancer du sein*, *Octobre rose 2014*, *Cancer du sein - breast cancer, brustkrebs*

[8] Institut National de Cancer - http://www.e-cancer.fr/Dictionnaire/

[9] http://aims.fao.org/fr/agrovoc

On peut lui associer au plus un label préféré (*skos:prefLabel*), c'est-à-dire la dénomination privilégiée du concept. D'autres termes peuvent être associés au concept comme variantes valides (*skos:altLabel*) ou variantes existantes mais déconseillées (*skos:hiddenLabel*).

Ce modèle initial ne suffit pas pour conserver les méta-données relatives au processus d'extraction de chaque relation patient/médecin, à savoir le poids de la relation, la méthode ayant généré la relation et enfin son type. Nous avons donc étendu notre usage de SKOS pour intégrer la provenance de la relation, en particulier à l'aide du vocabulaire PROV[10] qui est une recommandation du W3C pour représenter les informations de provenance. Le modèle final obtenu se présente comme décrit sur la figure 1.

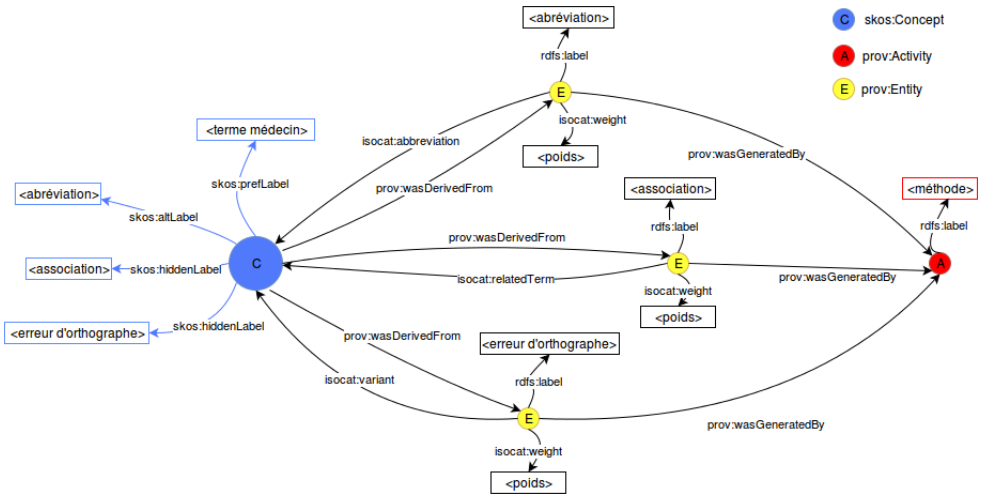


Figure 1: Modèle de representation des relations patient/médecin en SKOS+PROV dans MuEVo

Chaque *skos:Concept* (représenté en bleu sur la figure) est une représentation formelle de toutes les relations trouvées pour un "terme médecin" donné. Pour un *skos:Concept* donné, l'identifiant implicite est le "terme médecin" qui le décrit. Il doit donc être unique. On l'assigne alors au champ *skos:prefLabel*. Chaque mesure mise en jeu est représentée par une *prov:Activity* (en rouge sur la figure). Par souci de lisibilité, une seule méthode est mentionnée sur la figure mais plusieurs peuvent être utilisées. Chaque relation reliant le "terme médecin" du concept à un "terme patient" est représentée via les labels standards SKOS (*skos:altLabel* ou *skos:hiddenLabel*). En complément, nous conservons les informations de provenance à l'aide d'une entité RDF de type *prov:Entity* reliée au concept par un label ISOcat qui sert à préciser le type de la relation. La fonction de détermination des labels SKOS et ISOcat est donnée par le tableau 2. Le poids de la relation est également stocké dans la *prov:Entity* correspondante à l'aide du label *isocat:weight*. Chaque entité modélisant une relation avec le "terme médecin" du concept est stockée dans le *skos:Concept* associé au "terme médecin" à l'aide d'une information de provenance *prov:wasDerivedFrom*.

Après formalisation des relations patient/médecin en SKOS, nous souhaitons aligner *MuEVo* à des terminologies de référence.

[10]https://www.w3.org/TR/prov-dm/

Type de la relation	Label SKOS	Label ISOcat
abréviation	*skos:altLabel*	*isocat:abbreviation*
erreur d'orthographe	*skos:hiddenLabel*	*isocat:variant*
association	*skos:hiddenLabel*	*isocat:relatedTerm*

Tableau 2 : Fonction d'attribution des labels SKOS et ISOcat

3 Alignement du vocabulaire médecin

BioPortal (Noy *et al.*, 2009) est un serveur de terminologies biomédicales. Dans le cadre du projet SIFR[11], une instance de BioPortal[12] donne accès à une version *en français* des principales terminologies du domaine biomédical (Jonquet *et al.*, 2016). Via ce portail Web, un utilisateur peut partager une terminologie sur le serveur et la relier à celles déjà disponibles via des mappings étiquetés là encore à l'aide de SKOS. Nous avons donc chargé MuEVo dans SIFR BioPortal après l'avoir formalisé précédent et souhaitons maintenant relier nos concepts à ceux des terminologies biomédicales standards disponibles.

Le vocabulaire médecin initial, la liste de l'INCa, est une liste plate et de taille réduite. La création de ces liens (mappings) nous permettra de bénéficier de la connaissance plus large et structurée offerte par ces terminologies lors de l'usage explicite du vocabulaire patient/médecin pour indexer sémantiquement le contenu de forums par exemple.

Nous visons uniquement l'établissement de liens d'équivalence *skos:exactMatch* et de liens hiérarchiques : hyperonymie ou généralisation (*skos:broadMatch*) et hyponymie ou spécialisation (*skos:narrowMatch*). Pour nos premières expérimentations, nous nous sommes limités à trois terminologies cibles en français : MeSH, SNOMED et MedDRA. L'approche d'alignement adoptée s'articule en deux phases : un alignement direct et indirect.

3.1 Alignement direct

La phase d'alignement direct consiste à rechercher à l'aide de l'API REST[13] de BioPortal chaque terme de l'INCa dans notre vocabulaire. Si l'on retrouve exactement le même terme comme appellation préférée ou variante d'un concept d'une terminologie cible, alors on établit un lien d'équivalence, *skos:exactMatch*, entre le concept étudié et celui de la terminologie cible. Sur la figure 2, le terme *abdomen* est l'appellation préférée d'un concept d'une terminologie standard. Le concept *Abdomen* est alors relié. Le terme *cancer* apparaît comme variante du concept standard *Tumeurs* donc un lien *skos:exactMatch* est créé.

Pour les termes n'apparaissant comme label d'aucun concept des terminologies cibles, nous recherchons un alignement indirect.

[11] Semantic Indexing of French Biomedical Data Resources - http://www.lirmm.fr/sifr/

[12] http://bioportal.lirmm.fr/

[13] http://data.bioportal.lirmm.fr/documentation

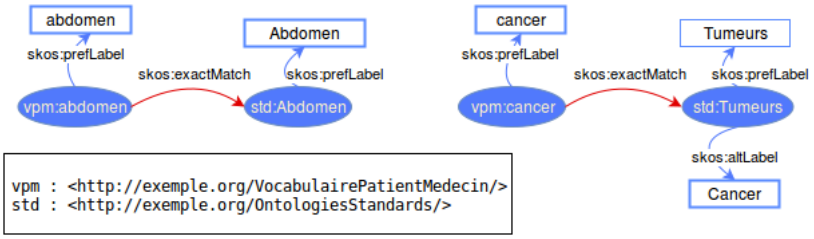


Figure 2: Exemples d'alignements directs

3.2 Alignement indirect

Nous faisons ici l'hypothèse qu'il existe des ressources plus généralistes intermédiaires entre la liste de l'INCa et les entrées des terminologies standards cibles. Ainsi, pour un "terme médecin" t_m donné de MuEVo, il s'agit d'utiliser des ressources externes, Wiktionary[14] (Meyer & Gurevych, 2012) dans notre cas, pour trouver des termes en lien avec t_m par une relation sémantique de type *synonyme, hyperonyme, hyponyme* et qui apparaissent eux comme labels dans les terminologies cibles. Le protocole adopté se décrit comme suit :

1. On recherche[15] le terme médecin dans Wiktionary. Si l'entrée existe alors on récupère l'ensemble des synonymes, hyperonymes et hyponymes

2. Pour chaque terme *t* de la liste ainsi constituée, une recherche parmi les labels des terminologies cibles à l'aide l'API de BioPortal est effectuée

3. En cas de succès, on définit les mappings suivants entre notre concept initial $C_{initial}$ et le concept C_{cible} de la terminologie cible retourné par l'API de recherche : si *t* était un synonyme : $C_{initial}$ *skos:exactMatch* C_{cible}; si *t* était un hyperonyme : $C_{initial}$ *skos:broadMatch* C_{cible}; si *t* était un hyponyme : $C_{initial}$ *skos:narrowMatch* C_{cible}.

Par exemple (voir figure 3), pour le terme *cure*, un synonyme est *traitement*; *oncologue* a pour hyperonyme *médecin spécialiste* et un hyponyme de *atome* est *ion*.

4 Résultats

Le vocabulaire MuEVo est consultable[16] sur SIFR BioPortal. Les résultats de l'alignement des 64 termes médecin du vocabulaire étudié sont résumés dans le tableau 3. Ces alignements ont été réalisés via un programme que nous avons écrit pour automatiser le processus. Les trois terminologies cibles choisies couvrent 84,38% du vocabulaire médecin : MeSH (70,31%), SNOMED (51,56%) et MedDRA (37,5%). Parmi les 10 termes manquants, 3 ont pu être alignés avec succès grâce aux hyponymes extraits de Wiktionary. Pour les 7 restants, l'alignement est manuel.

[14]http://www.wiktionary.org

[15]Nous automatisons la recherche en utilisant l'API JWKTL (Zesch *et al.*, 2008) après l'avoir adapté pour le français

[16]MuEVo - http://bioportal.lirmm.fr/ontologies/MUEVO

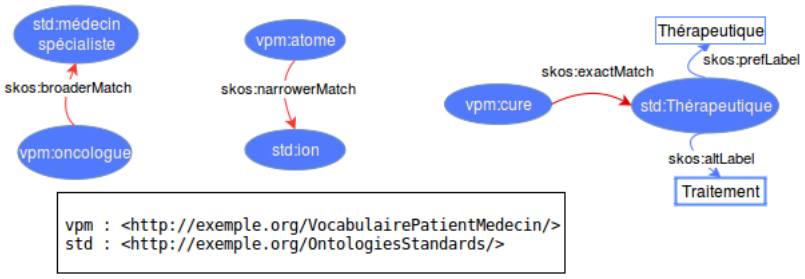


Figure 3: Exemple d'alignement indirect pour les termes oncologue, atome et cure

	Nombre	**Exemples**
1A : Singulier	51	abdomen -> Abdomen (MeSH)
1B : Pluriel	17	glucide -> Glucides (MeSH)
1A+1B	54	
2 : Hyponymes	3	atome -> ion (SNOMED)

Tableau 3 : Résultats obtenus automatiquement pour 64 termes en entrée de la phase d'alignement direct (1A, 1B) et 10 termes à la phase d'alignement indirect 2

5 Conclusions et perspectives

Dans cet article, nous avons proposé une méthode de formalisation d'un vocabulaire patient/médecin en terminologie au format SKOS ainsi que des pistes pour aligner le vocabulaire médecin correspondant aux terminologies de référence existantes. Une telle ressource peut être utilisée pour rendre des productions médicales (dossiers médicaux par exemple) plus compréhensibles aux patients (Zeng & Tse, 2006) ou pour de l'indexation multi-expertise (Soualmia *et al.*, 2003). Dans la suite de nos travaux, nous envisageons d'explorer trois points : la structuration interne de MuEVo à l'aide des relations sémantiques extraites de définitions (Medelyan *et al.*, 2009), l'acquisition de nouvelles relations patient/médecin en utilisant le métathésaurus UMLS[17] (Keselman *et al.*, 2008), plus large que celui de l'INCa et enfin l'exploitation de la ressource pour des tâches de classification supervisées et non supervisées exploitant la hiérarchie des terminologies auxquelles MuEVo est aligné (Wijewickrema *et al.*, 2015).

6 Remerciements

Ce travail est réalisé au sein du projet SIFR financé par le programme JCJC ANR-12-JS02-01001 et le projet "Comparison of longitudinal analysis models of the health-related quality of life in oncology" financé par l'IRESP.

[17]Unified Medical Language System - https://www.nlm.nih.gov/research/umls/

References

DELAVIGNE V. (2012). Peut-on «traduire» les mots des experts? un dictionnaire pour les patients atteints de cancer. *Dictionnaires et traduction*, p. 233–263.

JIANG L., YANG C. C. & LI J. (2013). Discovering consumer health expressions from consumer-contributed content. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, p. 164–174: Springer.

JONQUET C., ANNANE A., BOUARECH K., EMONET V. & MELZI S. (2016). SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In *16th Journées Francophones d'Informatique Médicale, JFIM'16*.

KESELMAN A., SMITH C. A., DIVITA G., KIM H., BROWNE A. C., LEROY G. & ZENG-TREITLER Q. (2008). Consumer health concepts that do not map to the umls: where do they fit? In *Journal of the American Medical Informatics Association*, volume 15, p. 496–505: Elsevier.

KOGAN S., ZENG Q., ASH N. & GREENES R. A. (2001). Problems and challenges in patient information retrieval: a descriptive study. In *Proceedings of the AMIA Symposium*, p. 329: American Medical Informatics Association.

MCCRAY A. T., LOANE R. F., BROWNE A. C. & BANGALORE A. K. (1999). Terminology issues in user access to web-based medical information. In *Proceedings of the AMIA Symposium*, p. 107: American Medical Informatics Association.

MEDELYAN O., MILNE D., LEGG C. & WITTEN I. H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, **67**(9), 716–754.

MEYER & GUREVYCH (2012). Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In *Granger, S. and Paquot, M., International Conference on Dublin Core and Metadata Applications*.

MILES A., MATTHEWS B., WILSON M. & BRICKLEY D. (2005). Skos core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, p. pp–3.

NOY N. F., SHAH N. H., WHETZEL P. L., DAI B., DORF M., GRIFFITH N., JONQUET C., RUBIN D. L., STOREY M.-A., CHUTE C. G. *et al.* (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. In *Nucleic acids research*, p. gkp440: Oxford Univ Press.

SOUALMIA L., DARMONI S. J., DOUYÈRE M. & THIRION B. (2003). Modelisation of consumer health information in a quality-controlled gateway. In *Studies in health technology and informatics*, p. 701–706: IOS Press; 1999.

TAPI NZALI M. D., BRINGAY S., LAVERGNE C., OPITZ T., AZÉ J. & MOLLEVI C. (2015). Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. In *Ingénierie des Connaissances*, p. 9–20.

WIJEWICKREMA C. M. *et al.* (2015). Impact of an ontology for automatic text classification. *Annals of Library and Information Studies (ALIS)*, **61**(4), 263–272.

ZARRO M. & LIN X. (2011). Using social tags and controlled vocabularies as filters for searching and browsing: A health science experiment. *Mountain View, CA*.

ZENG Q. T. & TSE T. (2006). Exploring and developing consumer health vocabularies. In *Journal of the American Medical Informatics Association*, volume 13, p. 24–29: Elsevier.

ZESCH T., MÜLLER C. & GUREVYCH I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, p. 1646–1652.

Interopérabilité sémantique dans le domaine du diagnostic *in vitro*

Mélissa Mary[1,2], Lina F. Soualmia[2,3] et Xavier Gansel[1]

[1] bioMérieux SA, Département Développement et Intégration,
38390 La Balme Les Grottes
{melissa.mary,xavier.gansel}@biomerieux.com
http://biomerieux.com
[2] LITIS EA 4108 et NormaSTIC CNRS 3638, Université de Normandie, 76000 Rouen
Lina.Soualmia@chu-rouen.fr
[3] LIMICS INSERM UMR_1142, Sorbonne Universités, 75000 Paris

Résumé : L'interopérabilité entre Systèmes d'Organisation de la Connaissance (SOC) est un des enjeux lié à l'émergence des dossiers médicaux partagés. Dans cet article nous proposons une évaluation de méthodes d'alignement de concepts issus du diagnostic *in vitro* (DIV).Les méthodes proposées reposent sur trois mesures de similarité syntaxique et un algorithme à base d'heuristiques. Les résultats que nous obtenons dans cette étude montrent que les métriques de similarité syntaxique ne se révèlent pas suffisamment probantes pour se voir appliquées de manière systématique au domaine des tests de laboratoire. En revanche, la qualité des alignements obtenus via un algorithme heuristique, filtré a posteriori en fonction d'une dimension sémantique, permettent de conforter les critères de performance que nous avions établis. Cet algorithme est notre piste privilégiée pour obtenir des alignements de qualité dans le domaine du DIV. Il est en cours d'amélioration par l'enrichissement, à la volée, d'informations syntaxiques et sémantiques.

Mots-clés : algorithme d'alignement, domaine de la santé, évaluation, interopérabilité sémantique, systèmes d'organisation de la connaissance.

1 Introduction

L'informatisation croissante des données médicales au sein de dossiers médicaux partagés a pour objectif d'améliorer la prise en charge du patient par l'établissement d'échanges d'informations interopérables entre acteurs et systèmes de la chaîne de soins (Fieschi, 2009; Macary, 2007; Stroetmann, 2009). Dans le domaine du diagnostic *in vitro* (DIV) deux Systèmes d'Organisation de la Connaissance (SOC) sont recommandés par les instances de standardisation nationales et internationales pour codifier les informations (Blumenthal, 2010; Stroetmann, 2009) : la terminologie LOINC®[1] (Logical Observation Identifiers Names and Codes) pour la description des tests, et l'ontologie SNOMED CT® (Systematized Nomenclature of MEDicine – Clinical Terms) pour coder les résultats. Une collaboration a récemment été mise en place (IHTSDO et Regenstrief Institute, 2013) afin d'aligner LOINC® et SNOMED CT®. L'objectif de cet alignement est d'améliorer l'agrégation de données de comptes rendus d'analyse dans les dossiers patients informatisés (Vreeman, 2015).

De plus en plus d'initiatives visent à réaliser des alignements entre SOC dans le domaine clinique. L'obtention des alignements est un enjeu majeur tant du fait de la volumétrie des données à aligner que de la qualité des alignements. De nombreuses méthodes, stratégies et métriques ont été développées permettant de réaliser des alignements notamment entre ontologies (Brahma & Refoufi, 2015; Euzenat & Shvaiko, 2013). Dans cet article, notre étude

1 http://loinc.org/

vise à évaluer trois métriques de similarité et un algorithme heuristique sur des données de DIV.

En tenant compte des caractéristiques des SOC et l'emploi de ces alignements, nous avons établi plusieurs critères d'évaluation pour discriminer ces méthodes. L'évaluation est possible grâce à des alignements existants entre la terminologie LOINC® et l'ontologie SNOMED CT® réalisés par des experts.

2 Matériel et Méthode

2.1 Matériel

Logical Observation Identifiers Names and Codes

La terminologie LOINC® a été construite en 1994 afin de standardiser la description des tests cliniques et de diagnostic *in vitro* (Sheide & Wilson, 2013). Un test codé en LOINC® se décompose en 6 parties (*Composant*, *Milieu*, *Technique, Méthode Échelles* et *Grandeur)*. Afin de compacter la description d'un test, les parties *Milieu*, *Échelles* et *Grandeurs* sont représentées par des codes mnémoniques : par exemple « *blood* » est encodé par « bld ». On retrouve également certains mots abrégés dans les parties *Composant* : « antigène » en « Ag » par exemple. Dans cette étude nous utilisons la version 2.5 de LOINC®, les parties et leur hiérarchie sont extraites de la version 2.5 de la base de données utilisée par l'application RELMA.

Systematized Nomenclature of MEDicine – Clinical Terms La SNOMED CT® est une ontologie dédiée à la représentation d'information du domaine clinique (Cornet & de Keizer, 2008). Les concepts sont organisés dans 19 axes et identifiés par un libellé unique qui se compose d'un terme spécifiant la sémantique du concept et d'un tag sémantique. Cette étude a été réalisée à l'aide de la version de SNOMED CT® datant de janvier 2015.

Alignement LOINC® SNOMED CT®

Les alignements entre LOINC® et SNOMED CT® utilisés dans cette étude sont le résultat d'une collaboration initiée en 2013 entre l'IHTSDO et le *Regenstrief Institute* (IHTSDO & Regenstrief Institute, 2013). L'alignement de LOINC® sur SNOMED CT® est réalisé à deux niveaux. Dans un premier temps les parties ont été alignées sur des concepts SNOMED CT®. Par la suite les tests LOINC® sont décrits par des définitions formelles en SNOMED CT®.

Dans cette étude nous utilisons la première version d'alignement publiée en septembre 2014, qui couvre 0,15% des tests LOINC® et 2,115 parties (5%). Dans cette étude nous utilisons les 2 177 alignements entre concepts SNOMED CT® et les parties LOINC®.

2.2 Méthodes

Les méthodes présentées ont été implémentées dans le langage R (R. Core Team, 2014).

2.2.1 Similarités syntaxiques

Damerau Levenshtein

La similarité de Damerau-Levenshtein (DL) (Damerau, 1964; Levenshtein, 1966) se calcule à partir de la distance d'édition du même nom fournit par le package stringdist (Van der Loo, 2014). Elle repose sur l'équation suivante :

$$sim_{DL}(t_1, t_2) = 1 - \frac{dist_{DL}(t_1, t_2)}{min(|t_1|, |t_2|)} \qquad (1)$$

Où t_1, t_2 représentent deux termes et $|t_1|$ $|t_2|$ leur longueur.

Stoilos

La similarité de Stoilos est une métrique développée spécifiquement pour les libellés de concepts d'ontologies (Stoilos et al., 2005). Cette similarité (équation 2) pondère positivement les termes ayant un préfixe commun (fonction *winkler*).

$$sim_{Stoilos}(t_1, t_2) = common(t_1, t_2) - diff(t_1, t_2) + winkler(t_1, t_2) \quad (2)$$

$$common(t_1, t_2) = \frac{2 * \sum_{i=1}^{n} |substring(t_1, t_2)|_i}{|t_1| + |t_2|} \quad (3)$$

$$diff(t_1, t_2) = \frac{diff_1 * diff_2}{p_{diff} + (1 - p_{diff})(|diff_1| + |diff_2| - |diff_1| * |diff_2|)} \quad (4)$$

$$winkler(t_1, t_2) = p_{winkler} * (1 - common(t_1, t_2)) * min(4, |common_{prefix}|) \quad (5)$$

Où *common(t_1, t_2)* représente la communauté entre les deux chaînes de caractères t_1 et t_2, (équation 3), *diff(t_1, t_2)* la différence entre t_1 et t_2 (équation 4) et *winkler(t_1, t_2)* permet l'amélioration du résultat en utilisant la méthode introduite par Winkler (Winkler, 1999) (équation 5).

WGram

La méthode WGram s'appuie sur la décomposition des termes en vecteur de mots (w) et produit pour un alignement deux scores de similarité (terme 1 vers terme 2 et terme2 vers terme1). La similarité d'un terme par rapport à un autre (équation 6) se calcule par le biais des similarités des mots qu'ils ont en commun.

$$sim_{WGram}(t_1 \rightarrow t_2) = \frac{\sum_{i=1}^{k} |w_1^i| * sim_{DL}(w_1^i, w_2)}{\sum_{j=1}^{N} |w_1^j|} \quad (6)$$

Où k représente le nombre de mots alignés entre les termes t_1 et t_2, N représente le nombre de mots (w^i_1) composant t_1 et w_2 un des mots composant le terme t_2.

2.2.2 Similarité sémantique

La similarité sémantique entre deux termes est calculée de manière indirecte grâce au Metathesaurus® développé par l'US National Library of Medicine (Fung & Bodenreider, 2005). Le Metathesaurus intègre les termes 195 ressources biomédicales (13 millions de termes) qui sont représentés par 3 millions de concepts (CUI). Le calcul de similarité sémantique s'effectue en 2 étapes. La première étape consiste à extraire les CUI correspondant aux termes. Nous avons développé un algorithme basé sur l'outil MetaMap (Aronson & Lang, 2010) qui permet de compiler pour chaque terme la liste des termes sémantiques candidats associée. Un terme sémantique candidat est composé de deux informations : (i) une liste de CUI et (ii) un indice de confiance de l'alignement entre le terme et son candidat. La seconde étape consiste à calculer la similarité sémantique entre les deux termes. Nous utilisons la métrique de hamming ($sim_{hamming}$). La similarité de hamming se calcule à partir de la liste des CUI des termes sémantiques candidats (t^i_{sem1}, t^j_{sem2}) (équation 7).

$$sim_{sem}(t_1, t_2) = max(sim_{hamming}(t^{'}_{sem1}, t^{'}_{sem2})) \quad (7)$$
$$t^{'}_{sem1} \in \lfloor t_{sem} \rfloor_1, t^{'}_{sem2} \in \lfloor t_{sem} \rfloor_j$$

Nous mesurons la confiance de cette similarité comme étant le minimum entre la confiance sémantique des termes sémantiques candidats (t^i_{sem1}, t^j_{sem2}) représentant respectivement t_1 et t_2.

$$conf_{sim.sem}(t^i_{sem1}, t^j_{sem2}) = min(conf_{sem}(t^i_{sem1}), conf_{sem}(t^j_{sem2})) \quad (8)$$

Paramètres utilisés pour MetaMap

Pour cette étude, nous considérons chacun des termes des deux SOC comme étant un texte et nous effectuons la recherche de concepts sur une version modifiée du Metathesaurus® (2014AB) qui exclut les ressources LOINC® et SNOMED CT®.

2.2.3 Alignement heuristique par la méthode des ancres

La méthode des ancres que nous proposons est une stratégie d'alignement heuristique qui s'inspire d'un algorithme développé pour résoudre les problématiques d'alignement d'ontologies volumineuses (Seddiqui & Aono, 2009).

Cette méthode part du postulat que les données au sein des deux ressources à comparer sont organisées de manière similaire. Le principe de la méthode consiste à produire un nouvel alignement par extension d'un alignement déjà existant (alignement initial). L'alignement initial permet de définir un ensemble d'ancres (dites ancres initiales) qui serviront à générer de manière récursive de nouvelles ancres (ou ancres générées). Soit t_1 et t_2 .les termes des deux ressources composant une ancre. L'algorithme réalise trois alignements distincts entre (i) les termes parents de t_1 et t_2 (ii) les termes enfants de t_1 et t_2 et (iii) les termes frères t_1 et t_2. Dans une seconde étape les alignements sont sélectionnés par un critère de seuil pour déterminer les nouvelles ancres.

Dans cette étude nous appliquons la méthode de DL pour générer les alignements et nous utilisons un seuil de similarité comme paramètre de filtre pour créer de nouvelles ancres.

2.3 Évaluation

L'alignement ($\mathcal{A}$) entre les SOC ($\mathcal{T}_1$et $\mathcal{T}_2$) est évalué grâce à un alignement de référence (noté $\mathcal{M}$) par les paramètres de précision (équation 9) et de rappel (équation 10).

$$precision(\mathcal{A}_{\mathcal{T}_1,\mathcal{T}_2}, \mathcal{M}_{\mathcal{T}_1,\mathcal{T}_2}) = \frac{|\mathcal{A}_{\mathcal{T}_1,\mathcal{T}_2} \cap \mathcal{M}_{\mathcal{T}_1,\mathcal{T}_2}|}{|\mathcal{A}_{\mathcal{T}_1,\mathcal{T}_2}|} \quad (9)$$

$$rappel(\mathcal{A}_{\mathcal{T}_1,\mathcal{T}_2}, \mathcal{M}_{\mathcal{T}_1,\mathcal{T}_2}) = \frac{|\mathcal{A}_{\mathcal{T}_1,\mathcal{T}_2} \cap \mathcal{M}_{\mathcal{T}_1,\mathcal{T}_2}|}{|\mathcal{M}_{\mathcal{T}_1,\mathcal{T}_2}|} \quad (10)$$

2.4 Normalisation des termes et filtre des alignements

Normalisation des termes

Nous avons défini une méthode de normalisation des termes par SOC. Dans les parties LOINC les éléments de ponctuations sont supprimé ou remplacés par des espaces. Les libellés des concepts SNOMED CT® sont normalisés au niveau de la ponctuation et le tag sémantique est supprimé.

Paramètres de filtre des alignements

Nous utilisons deux filtres pour étudier les alignements calculés à partir des similarités syntaxiques. Par défaut, un alignement entre deux SOC ($\boldsymbol{\mathcal{T}}_1$et $\boldsymbol{\mathcal{T}}_2$) est composé des meilleurs alignements par terme du SOC 1 et des meilleurs alignements par terme du SOC 2 (équation 11).

$$\mathcal{A}(\mathcal{T}_1, \mathcal{T}_2) = \{\mathcal{A}(t_x, t_y) | sim(t_x, t_y) = max(sim(t_x, \mathcal{T}_2)) \vee sim(t_x, t_y) = max(sim(\mathcal{T}_1, t_y))\} \quad (11)$$

Le filtre `BestForBoth` (équation 12) que nous proposons permet de sélectionner les alignements qui sont les meilleurs à la fois pour t_1 et pour t_2.

$$BestForBoth(\mathcal{T}_1, \mathcal{T}_2) = \{\mathcal{A}(t_x, t_y) | sim(t_x, t_y) = max(sim(t_x, \mathcal{T}_2)) = max(sim(\mathcal{T}_1, t_y))\} \quad (12)$$

Les résultats de ces deux filtres sont contextuels. Ils sont fonction à la fois de la métrique utilisée mais surtout des deux ensembles de termes utilisés pour l'alignement.

3 Résultats et discussion

L'objectif de cette évaluation est d'identifier les méthodes et métriques et les plus performantes pour l'alignement de termes spécifiques au domaine du diagnostic *in vitro*. Les SOC représentant des données de DIV ont trois caractéristiques majeures qui impactent directement les critères de performance requis pour un alignement. La première caractéristique, et la plus critique, concerne la qualité des ressources. Afin d'être mis en œuvre dans les systèmes, les alignements produits doivent être les plus précis possible, c'est-à-dire que la méthode doit maximiser le nombre de vrais alignements. La seconde caractéristique est relative à la quantité de données intégrée dans les SOC. Enfin, la troisième caractéristique concerne l'hétérogénéité des champs lexicaux en fonction du sous-domaine de connaissances représenté dans les SOC. Par exemple, les termes associés à la dénomination des microorganismes suivent une nomenclature stricte alors que le vocabulaire pour exprimer un milieu est plus riche en variation linguistique. Les métriques et méthodes sont évaluées sur deux critères :

- une précision maximale pour un nombre d'alignements cohérent avec l'ordre de grandeur du plus petit SOC aligné ;
- la robustesse des métriques pour des sous-domaines de connaissances dont l'expression des termes peut être plus ou moins permissive.

3.1 Évaluation des méthodes de similarité syntaxiques

3.1.1 Analyse des similarités syntaxiques

TABLE 1 *Résultat de l'évaluation des métriques de similarité syntaxique*.

Méthode	Normalisation des termes	Filtre	Nombre d'alignements	Précision	Rappel
DL	Oui	Non	4 589	0,26	0,54
	Oui	`Sim ≥ 0,8`	1 192	0,76	0,41
	Oui	`BestForBoth`	1 281	0,77	0,45
Stoilos (p_{diff} = 0.6 ; $p_{winkler}$ = 0.1)	Non	Non	3 132	0,50	0,72
	Non	`BestForBoth`	**1 356**	0,92	0,57
WGram	Non	Non	3 263	0,47	0,71
	Non	`BestForBoth`	2 322	0,63	0,67
WGram et Stoilos	Non	`BestForBoth pour la métrique Stoilos`	1 202	**0,95**	0,53

Le tableau 1 résume l'évaluation des alignements obtenus par les métriques de similarité syntaxique. Tout d'abord on observe que les rappels associés aux méthodes syntaxiques sont faibles (40 à 70%). L'analyse du jeu de données initial montre que 12% des parties LOINC sont représentées par des codes mnémoniques ce qui explique en partie le faible rappel.

Après application du filtre `BestForBoth` sur les alignements DL et Stoilos, on peut constater une augmentation significative de la précision (+50%) au détriment du nombre de vrais alignements retrouvés (diminution de 10-15% du rappel).

On observe également que la meilleure précision (0,95) est obtenue par l'intersection des résultats d'alignement WGram avec Stoilos.

3.1.2 Comparaison des performances du filtre BestForBoth avec le filtre par seuil

Les performances des méthodes de similarité sont généralement étudiées avec un paramètre de seuil comme filtre d'alignement. Dans cette étude nous avons décidé de ne pas investiguer les performances associées à ce type de filtre. Ce choix est motivé par une question majeure : Comment déterminer sans a priori le seuil de filtrage ?

***TABLE 2** Exemple d'alignement entre LOINC® et SNOMED CT ® obtenu avec la méthode DL*

Partie LOINC®		Concept SNOMED CT		Similarité DL	Vrai Alignement ?	BestForBoth	Sous domaine
LP14078-7	Babesia bovis	43574002	Babesia ovis	0,92	Non	Non	Organisme
LP14419-3	Leukocytes	52501007	Leukocyte	0,89	Oui	Oui	Cellule
LP7057-05	Bld	87612001	Blood	0,33	Oui	Oui	Milieu

La variabilité de la représentation terminologique des concepts dépend fortement du sous-domaine étudié. Certaines dimensions LOINC par exemple sont principalement représentées par des codes mnémoniques, ou contiennent des abréviations ce qui entraîne une forte variabilité lexicale induisant des similarités assez faibles (tableau 2 exemple « blood »). Le tableau 2 illustre cette problématique avec trois exemples retrouvés fréquemment dans les alignements. L'emploi d'un seuil de similarité pour filtrer les alignements sur les similarités syntaxiques doit être paramétré en fonction de la variabilité lexicale du sous-domaine étudié.

3.2 Évaluation de la stratégie des ancres

3.2.1 Évaluation des termes sémantiques candidats

Avant d'utiliser le calcul de similarité sémantique pour filtrer *a posteriori* les données obtenues par la méthode des ancres, nous avons vérifié la cohérence des termes sémantiques candidats extraits de MetaMap.

On observe que 24% des parties LOINC® (688) et 17% des concepts SNOMED CT® impliqués dans l'alignement n'ont aucun terme sémantique candidat. On observe que les proportions de parties sans terme sémantique sont plus élevées dans les dimensions *Milieu* (60%), *Unité* (83%) et *Méthode* (40%) qui sont représentés par des termes abrégés ou des codes mnémoniques. Nous pouvons en conclure que l'utilisation des termes sémantiques candidats ne permet pas de résoudre les problématiques liées aux codes mnémoniques pour désigner un concept. Cependant on observe que plus de 50% des termes sont représentés par un unique terme sémantique candidat, ce qui nous conforte dans l'idée de que l'UMLS® est une ressource suffisamment précise pour calculer la similarité sémantique entre deux termes du DIV.

On observe également que les termes sans candidat sémantique sont impliqués dans 45% des alignements initiale (945), ce qui peut expliquer le faible rappel obtenu par l'application du filtre sémantique sur les ancres initiales (voir tableau 3).

3.2.2 Performances de la stratégie des ancres et du filtre sémantique *a posteriori*

Pour générer des alignements avec la méthode des ancres nous avons utilisé comme jeu de données initial les alignements créés par la collaboration entre le Regenstrief et l'IHTSDO (tableau 3 ancres initiales). Nous avons choisi de réaliser le calcul des nouvelles ancres par la méthode DL en appliquant un seuil de 0,8 sur la similarité, ce seuil permet d'obtenir des performances similaires au filtre `BestForBoth` sur le jeu de données initial (voir tableau 1). Ici nous n'appliquons pas le filtre `BestForBoth` pendant la génération des ancres parce que le nombre de termes alignés par chaque itération de l'algorithme heuristique est trop petit pour garantir l'efficacité d'un tel filtre. On observe que les ancres générées ont une précision de l'ordre de 30% avant d'être filtrées avec des seuil de similarité sémantique. L'application *a posteriori* des filtres de similarité sémantique ($sim_{Semantique}$) et `BestForBoth` permettent de doubler la précision de l'alignement sur les ancres générées. On observe également que la précision du filtre sémantique augmente de 10% si l'on tient compte de la confiance sémantique de l'alignement entre les termes et leurs termes sémantiques candidats. Les meilleures performances sont obtenues par combinaison des informations sémantiques et syntaxique. En effet on observe que les filtres (i) $(sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800) \vee sim_{DL} = 1$

et (ii) $sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800 \wedge BestForBoth$ permettent d'obtenir 80% de précision pour 80% de rappel.

TABLE 3 *Résultats de l'évaluation de la méthode des ancres et de la méthode de filtrage basé sur la similarité sémantique. Les ancres initiales correspondent aux alignements proposés par le Regenstrief et l'IHTSDO. Pour les ancres générées la précision et le rappel sont calculés à partir de la curation manuelle des données brutes (1ère ligne). Les champs marqués avec un NA dans le tableau sont des valeurs non informatives. (1) Aucun rappel ne peut être calculé car le alignement de référence est obtenu par curation à partir de ces données. (2) L'alignement est un sous-ensemble du jeu de données initial (IHTSDO et Regenstrief Institute, 2013), la précision est donc de 1.*

Méthodes	Nombre d'alignements	Paramètres de filtres	Précision	Rappel
Ancres DL à 0,80	1 833 ancres générés	Non	0.33	NA (1)
		BestForBoth	0.58	**0,97**
		$sim_{Semantique} \geq 0.5$	0,65	0,85
		$sim_{Semantique} = 1$	0,69	0,85
		$sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800$	0,81	0,83
		$(sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800) \vee sim_{DL} = 1$	**0,82**	0,87
	2 177 ancres initiales	$(sim_{Semantique} = 1 \wedge confiance_{Semantique} > 800)$	NA (2)	**0,37**

4 Conclusion

Lors de cette étude nous avons cherché à identifier les méthodes les plus appropriées pour aligner des données du DIV. Nous avons tout d'abord étudié les métriques de similarité syntaxique. Cette étude nous a permis de montrer qu'un filtre de sélection contextuel des meilleurs alignements (BestForBoth) permet d'améliorer la précision sur les métriques DL et Stoilos. L'utilisation de filtres basés sur un seuil de similarité nécessite un paramétrage dépendant des domaines de connaissances à aligner. Nous envisageons de définir cette valeur seuil sans *a priori* en utilisant par exemple des statistiques bayésiennes ou des analyses de variabilité intra SOC. Nous avons également montré que la combinaison de métriques syntaxiques permettait d'augmenter les performances.

Nous avons par la suite étudié les résultats issus d'un algorithme heuristique, qui couplé avec un filtre sémantique *a posteriori* permet d'améliorer la précision et le rappel. Pour compléter l'étude de la méthode des ancres, nous envisageons d'implémenter l'algorithme WGram pour générer des ancres, ainsi que d'intégrer les filtres de dimension sémantique pendant la recherche. Nous espérons ainsi obtenir plus de résultats d'alignement tout en conservant des performances acceptables (0,80 en précision et rappel). La dimension sémantique est une méthode dépendant d'une ressource externe et est fonction des termes en entrée. Nous avons montré que le Metathesaurus®, bien qu'intégrant près de 200 ressources ne permettait pas de garantir l'identification d'un concept UMLS pour l'ensemble des parties LOINC®, ou concepts SNOMED CT®. Pour minimiser l'impact de l'incomplétude de cette ressource, nous préconisons donc l'utilisation des paramètres sémantiques en complément de métriques syntaxiques.

L'un des enjeux dans l'alignement entre LOINC® et SNOMED CT®, ou de manière plus générale entre les SOC concerne l'alignement de termes abrégés ou sous forme mnémonique. La métrique sémantique que nous avons développée à partir du Metathesaurus ne permet pas de résoudre cette problématique. Nous envisageons cependant de recréer ou utiliser un dictionnaire d'abréviation de type Allie (Yamamoto, Yamaguchi, Bono, & Takagi, 2011).

Références :

ARONSON, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236.

BLUMENTHAL, D. (2010). Launching HITECH. *New England Journal of Medicine*, *362*(5), 382‑385. doi:10.1056/NEJMp0912825

BRAHMA, B., & REFOUFI, A. (2015). *Ontology Matching Algorithms*. Communication présentée au Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication (p. 89:1–89:5), New York, NY, USA : ACM. doi:10.1145/2816839.2816928

CORNET, R., & DE KEIZER, N. (2008). Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, *8*(Suppl 1), S2. doi:10.1186/1472-6947-8-S1-S2

DAMERAU, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

EUZENAT, J., & SHVAIKO, P. (2013). *Ontology matching*. Springer-Verlag.

FIESCHI, M. (2009). *La gouvernance de l'interopérabilité sémantique est au coeur du développement des systèmes d'information en santé* (rapport public Publication no).

IHTSDO & REGENSTRIEF INSTITUTE. (juillet 2013). Regenstrief and the IHTSDO are working together to link LOINC and SNOMED CT.

LEVENSHTEIN, V. I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals*. Communication présentée au Soviet physics doklady (vol. 10, p. 707–710).

MACARY, F. (2007). IHDE, CDA et LOINC : des composants d'interopérabilité au service du partage des résultats de biologie médicale. *Spectra biologie*, *26*(158), 51‑57.

R. CORE TEAM. (2014). *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing, Vienna, Austria*. *2013*. ISBN 3-900051-07-0.

SEDDIQUI, M. H., & AONO, M. (2009). An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(4), 344‑356. doi:10.1016/j.websem.2009.09.001

SHEIDE, A., & WILSON, P. S. (2013). Reading up on LOINC. *Journal of AHIMA/American Health Information Management Association*, *84*(4), 58–60.

STOILOS, G., STAMOU, G., & KOLLIAS, S. (2005). A String Metric for Ontology Alignment. Dans Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen (dir.), *The Semantic Web – ISWC 2005* (p. 624‑637). Springer Berlin Heidelberg.

STROETMANN, V. (2009). *Semantic Interoperability for Better Health and Safer Healthcare*. European Communities.

VAN DER LOO, M. P. (2014). The stringdist package for approximate string matching. *The R*.

VREEMAN, D. (7 novembre 2015). Guidelines for using LOINC and SNOMED CT Together. *Daniel Vreeman*. Repéré à https://danielvreeman.com/guidelines-for-using-loinc-and-snomed-ct-together-without-overlap/

WINKLER, W. E. (1999). *The state of record linkage and current research problems*. Communication présentée au Statistical Research Division, US Census Bureau, Citeseer.

YAMAMOTO, Y., YAMAGUCHI, A., BONO, H., & TAKAGI, T. (2011). Allie: a database and a search service of abbreviations and long forms. *Database: The Journal of Biological Databases and Curation*, *2011*. doi:10.1093/database/bar013

Extraction d'associations d'EIM à partir de dossiers patients : expérimentation avec les structures de patrons et les ontologies

Gabin Personeni[1], Marie-Dominique Devignes[1], Michel Dumontier[2], Malika Smaïl-Tabbone[1] et Adrien Coulet[1]

[1] LORIA (CNRS, Inria NGE, Université de Lorraine), Vandœuvre-lès-Nancy, F-54506, France
[2] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, Etats-Unis

Résumé : Les Dossiers Médicaux Electroniques (DME) constituent une ressource de grand intérêt pour étudier les Evènements Indésirables Médicamenteux (EIM). Nous proposons ici de fouiller les DME pour identifier des EIM fréquemment associés dans des sous-groupes de patients. Les EIM ayant des manifestations complexes, nous utilisons l'analyse formelle de concepts et ses structures de patrons, un cadre mathématique permettant la généralisation, en exploitant les connaissances du domaine médical formalisées dans des ontologies. Les résultats obtenus dans trois expériences montrent que cette approche est flexible et permet d'extraire des règles d'association à divers niveaux de généralisation.

1 Introduction

Les Evènements Indésirables Médicamenteux (EIM) apparaissent inégalement dans différents sous-groupes de patients. Leurs causes sont multiples: génétiques, métaboliques, interactions de médicaments. Des études précédentes ont montré que les EIM pouvaient être détectés et étudiés en fouillant des Dossiers Médicaux Electroniques (DME) (LePendu, et al., 2013). Nous souhaitons explorer les DME pour révéler qu'un groupe de patients sensible aux effets secondaires d'un médicament est également sensible aux effets secondaires d'un autre. Pour cela, nous proposons une méthode pour identifier des EIM fréquemment associés dans des sous-groupes de patients. Les manifestations des EIM étant variables et complexes, nous utilisons une extension de l'Analyse Formelle de Concepts (AFC) : les structures de patrons (Ganter & Kuznetsov, 2001), associée à des ontologies pour permettre la généralisation des EIM extraits des DME. Nous avons utilisé un jeu de DME de patients diagnostiqués avec le Lupus Erythémateux Disséminé (LED), une maladie auto-immune. Ces patients sont souvent sujets aux EIM, de par les traitements pour le LED et les maladies opportunistes qui l'accompagnent (Vasudevan & Ginzler, 2009). Ces DME ont été extraits de STRIDE, l'entrepôt de données cliniques de l'Hôpital universitaire de Stanford (Lowe, et al., 2009).

2 Matériel et méthodes

2.1 Corpus de données

Notre corpus de données est un ensemble de 6 869 DME, de patients anonymisés, diagnostiqués avec le LED. Ce corpus documente 451 000 visites à l'hôpital, avec leur dates relatives, diagnostics encodés en codes ICD9-CM (International Classification of Diseases, Clinical Modification) et prescriptions sous la forme de listes d'ingrédients, representés par leurs identifiants RxNorm. Aussi, nous employons le terme « médicament » pour désigner un ingrédient actif plutôt que le médicament sous sa forme commerciale.

Afin d'identifier des EIM fréquemment co-occurrents, nous devons d'abord extraire ces EIM des DME, puis sélectionner les patients qui en présentent au moins deux. Nous établissons d'abord, pour chaque patient, une liste d'EIM candidats. Pour deux visites consécutives dans le DME, nous extrayions l'ensemble des médicaments prescrits (une prescription) dans la première visite D_i, et les phénotypes P_i diagnostiqués durant la seconde. L'intervalle entre deux visites doit être d'au plus 14 jours : il est raisonnable de supposer qu'un effet secondaire se manifeste dans un court délai après la prescription. De plus, nous avons pu observer qu'augmenter cet intervalle maximal n'augmente pas le nombre de patients retenus dans notre corpus, comme illustré en Table 1.

Un EIM candidat C_i est donc un couple d'ensembles $C_i = (D_i, P_i)$. Nous ne conservons dans P_i que les phénotypes listés dans SIDER 4.1 comme effets secondaires d'un médicament de D_i. SIDER est une base de données d'indications et d'effet secondaires de médicaments (Kuhn, et al., 2016). Nous excluons ensuite les candidats où P_i est vide. Nous excluons aussi un EIM candidats (D_1, P_1) si il existe un autre EIM candidat (D_2, P_2) pour le même patient tel que $D_1 \subseteq D_2$: en effet, si une prescription est répétée pour un patient, cela indique qu'elle n'a pas été jugée dangereuse pour lui.

TABLE 1 – *Nombre de patients avec au moins 2 EIM sélectionnés, et nombre d'EIM pour ces patients, pour différents intervalles maximum entre deux visites*.

Intervalle (jours)	**1**	**2**	**6**	**10**	**14**	**18**	**22**	**26**	**30**
|Patients|	434	461	498	526	**548**	555	558	564	576
|EIM|	2 396	2 587	2 902	3 110	**3 286**	3 388	3 454	3 501	3 621

Par ce processus de sélection, nous obtenons un corpus de 3 286 EIM provenant de 548 patients présentant au moins 2 EIM. La Table 2 présente des exemples d'EIM qui pourraient être extraits de DME, qui serviront à illustrer les expériences présentées dans cet article.

TABLE 2 – *Exemple de corpus avec 3 patients ayant presenté 2 EIM chacun*

Patients	**EIM**
P1	({prednisone}, {ICD 599.8}) ; ({acetaminophen}, {ICD 599.9})
P2	({prednisone}, {ICD 599.8}) ; ({prednisone}, {ICD 719.4})
P3	({prednisone, acetaminophen}, {ICD 599.9}); ({acetaminophen}, {ICD 719.4})

2.2 Analyse Formelle de Concepts et structures de patrons

L'AFC (Ganter & Wille, 1999) est un cadre mathématique pour l'organisation d'un ensemble de données dans un treillis de concepts formels, *i.e.*, une structure hiérarchique où un concept représente un ensemble d'objets partageant des propriétés. Il permet l'extraction de règles d'associations.

En AFC, les données sont un ensemble d'objets, chacun décrit par un ensemble d'attributs binaires. Les structures de patrons généralisent l'AFC pour l'appliquer à des objets munis d'une description de nature complexe (pas forcément binaire), par exemple : des ensembles, graphes, intervalles, annotations par des classes d'une ontologie (Ganter & Kuznetsov, 2001; Coulet, et al., 2013). Une structure de patron est un triplet $(G, (\mathcal{D}, \sqcap), \delta)$ où *(i)* G est un ensemble d'objets, *(ii)* $\mathcal{D}$ est un ensemble de descriptions, *(iii)* δ est une fonction qui associe un objet à sa description, *(iv)* $\sqcap$ est un opérateur définissant un ordre partiel $\leq_\sqcap$ sur les éléments de $\mathcal{D}$, appelé *INF*, tel que $X \sqcap Y$ est la description la plus spécifique qui est plus générale que X et Y . Ainsi $X \leq_\sqcap Y$, signifiant que Y est plus spécifique que X est équivalent à $X \sqcap Y = X$. Cet opérateur permet la généralisation des descriptions d'objets. Ceci sera illustré pour les opérateurs que nous définirons en Section 3.

Dans les structures de patrons, l'opérateur $.^\square$ définit une connexion de Galois entre des ensembles d'objets et des descriptions, tel que :

$$A^\square = \sqcap_{g \in A}\, \delta(g) \quad \text{pour tout ensemble d'objets } A \subseteq G$$
$$d^\square = \{g \in G \mid d \leq_\sqcap \delta(g)\} \quad \text{pour toute description } d \in \mathcal{D}$$

Un concept de patrons est une paire (A, d) vérifiant $A^\square = d$ et $d^\square = A$. Dans notre étude, G est un ensemble de patients, à qui on associe par δ la description de leurs EIM dans $\mathcal{D}$. La Section 3 décrit trois expériences utilisant toutes les structures de patrons, mais avec une représentation différente des EIM et ainsi une structure de patrons $(G, (\mathcal{D}, \sqcap), \delta)$ différente.

2.3 Ontologies médicales

Nous utilisions deux ontologies biomédicales : ICD9-CM qui décrit des classes de phénotypes et l'Anatomical Therapeutic Chemical Classification System (ATC) qui décrit des classes de médicaments. Nous considérons seulement la seule hiérarchie de classes de ces

ontologies afin de généraliser la description des phénotypes et des prescriptions extraits des DME. De plus, nous utilisons uniquement les trois niveaux les plus spécifiques d'ATC : sous-groupes pharmacologiques, sous-groupes chimiques, substances chimiques.

3 Expériences

Dans cette Section, nous décrivons trois expériences utilisant les structures de patrons pour extraire des règles d'association entre des EIM. Chaque expérience utilise une représentation différente des EIM de chaque patient, chacune de ces représentations faisant un plus grand usage des ontologies.

3.1 Première expérience avec les structures de patrons

On définit ici la structure de patrons $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$ où les objets de G sont des patients, et les descriptions dans $\mathcal{D}_1$ sont des vecteurs de *sous-descriptions*. Chaque sous-description est un ensemble d'ensembles de médicaments, *i.e.*, un ensemble de prescriptions, associé à une classe de phénotypes du premier niveau d'ICD.

TABLE 3 – *Exemple de descriptions de patients pour la structure de patrons* $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$, *avec deux classes ICD de premier niveau : maladies du système urogénital (580-629), et maladies du système musculo-squelettal system et des tissues conjonctifs (710-739).*

	ICD 580-629	**ICD 710-739**
Patient P1	{{prednisone}, {acetaminophen}}	∅
Patient P2	{{prednisone}}	{{prednisone}}
Patient P3	{{prednisone, acetaminophen}}	{{acetaminophen}}

Par exemple, en considérant uniquement les deux classes ICD de la Table 3, les deux sous-descriptions associées au patient P1 sont :

$$\delta_{1,\,\text{ICD 580-629}}(\text{P1}) = \{\{\text{prednisone}\}, \{\text{acetaminophen}\}\} \quad \text{et} \quad \delta_{1,\,\text{ICD 710-739}}(\text{P1}) = \emptyset$$

Les sous-descriptions sont associées à un des classes de premier niveau d'ICD pour représenter des EIM : le patient présente un phénotype de cette classe après avoir été prescrit l'un des ensembles de médicaments dans cette sous-description. Nous définissons les sous-descriptions comme des ensembles de prescriptions, où aucune prescription n'est comparable à une autre par l'ordre partiel $\subseteq$. Nous définissons alors l'opérateur $\sqcap_1$ tel que, pour deux descriptions X et Y de $\mathcal{D}_1$ quelconques :

$$X \sqcap_1 Y = \texttt{max}(\subseteq, \{x \cap y \mid (x, y) \in X \times Y\})$$

où $\texttt{max}(\leq_i, S)$ est l'unique sous-ensemble des éléments maximaux de l'ensemble S pour l'ordre partiel $\leq_i$. Formellement, on définit $\texttt{max}(\leq_i, S) = \{s \mid \nexists x.\ (s \leq_i x)\}$. Dans le cas présent, `max` ne conserve dans la description que les prescriptions les plus spécifiques. Par exemple, en considérant 4 médicaments d_1, d_2, d_3, et d_4 :

$$\begin{aligned}
&\{\{d_1, d_2, d_3\}\} \sqcap_1 \{\{d_1, d_2\}, \{d_2, d_4\}\} \\
&= \texttt{max}\big(\subseteq, \{\{d_1, d_2, d_3\} \cap \{d_1, d_2\}, \{d_1, d_2, d_3\} \cap \{d_2, d_4\}\}\big) \\
&= \texttt{max}\big(\subseteq, \{\{d_1, d_2\}, \{d_2\}\}\big) \\
&= \{\{d_1, d_2\}\}
\end{aligned}$$

On ne conserve que $\{d_1, d_2\}$ puisque $\{d_2\} \subseteq \{d_1, d_2\}$ et $\{d_1, d_2\}$ est l'unique élément maximal par $\subseteq$. En effet la sémantique de $\{d_2\}$ – une prescription qui contient le médicament d_2 – est plus générale que la sémantique de $\{d_1, d_2\}$ – une prescription qui contient les médicaments d_1 et d_2.

A chaque patient est associée une sous-description pour chaque classe de premier niveau d'ICD, l'opérateur défini pour une seule sous-description peut être généralisé à un vecteur de sous-descriptions :

$$\delta_1(\mathrm{P1}) \sqcap_1 \delta_1(\mathrm{P2}) = \langle \delta_{1,1}(\mathrm{P1}), \ldots, \delta_{1,n}(\mathrm{P1}) \rangle \sqcap_1 \langle \delta_{1,1}(\mathrm{P2}), \ldots, \delta_{1,n}(\mathrm{P2}) \rangle$$
$$= \langle \delta_{1,1}(\mathrm{P1}) \sqcap_1 \delta_{1,1}(\mathrm{P2}), \ldots, \delta_{1,n}(\mathrm{P1}) \sqcap_1 \delta_{1,n}(\mathrm{P2}) \rangle$$

La Figure 1 présente le semi-treillis associé à la structure de patron et les données de la Table 3. Cet exemple illustre qu'avec cette structure de patron les informations sur les EIM sont perdues lors de la généralisation.

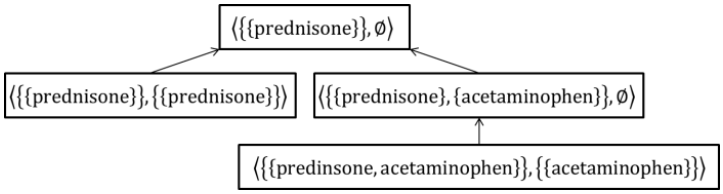


FIGURE 1 – Représentation des données de la Table 3 dans un semi-treillis construit à partir de la structure de patrons$(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$, où les flèches représentent l'ordre partiel $\leq_{\sqcap_1}$ défini par $\sqcap_1$.

3.2 Extension de la structure de patrons avec une ontologie de médicaments

Nous proposons de représenter les médicaments avec les termes de l'ontologie ATC afin de trouver des associations entre des EIM concernant des classes de médicaments plutôt qu'un simple médicament. Pour cela, nous devons étendre la structure de patrons décrite précédemment pour considérer cette ontologie. Chaque médicament est alors identifié par sa ou ses classes ATC, comme illustré en Table 4.

TABLE 4 – *Exemple de descriptions de patients pour la structure de patrons* $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$. H02AA03 : *desoxycortone*, H02AB07 : *prednisone*, N02BE01 : *acetaminophen*.

	ICD 580-629	**ICD 710-739**
Patient P1	$\{\{\mathrm{H02AB07}\}, \{\mathrm{N02BE01}\}\}$	$\emptyset$
Patient P2	$\{\{\mathrm{H02AB07}\}\}$	$\{\{\mathrm{H02AB07}\}\}$
Patient P3	$\{\{\mathrm{H02AB07}, \mathrm{N02BE01}\}\}$	$\{\{\mathrm{N02BE01}\}\}$
Patient P4	$\{\{\mathrm{H02AA03}\}\}$	$\emptyset$

On définit la seconde structure de patrons $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$, où les descriptions de $\mathcal{D}_2$ sont des ensembles de prescriptions dont les médicaments sont représentés par des classes ATC. Pour définir $\sqcap_2$, il nous faut comparer des ensembles de classes d'une ontologie $\mathcal{O}$, et donc définir un opérateur intermédiaire $\sqcap_{\mathcal{O}}$, tel que, pour x et y deux ensembles de classes de $\mathcal{O}$:

$$x \sqcap_{\mathcal{O}} y = \mathtt{max}\left(\sqsubseteq, \left\{LCA(c_x, c_y) \mid (c_x, c_y) \in x \times y\right\}\right)$$

où $LCA(c_x, c_y)$ est l'ancêtre commun le plus spécifique de c_x et c_y dans $\mathcal{O}$, et $\sqsubseteq$ est l'ordre partiel défini par la hiérarchie de classes de $\mathcal{O}$. Ainsi, $\mathtt{max}(\sqsubseteq, S)$ est le sous-ensemble des classes de S les plus spécifique, et $x \sqcap_{\mathcal{O}} y$ est l'ensemble des ancêtres communs les plus spécifiques des classes dans x et y. A partir de $\sqcap_{\mathcal{O}}$ nous définissons l'ordre partiel $\leq_{\mathcal{O}}$ comparant deux ensembles de classes de $\mathcal{O}$, tel que $x \leq_{\mathcal{O}} y$ signifie que y est un ensemble de classes plus spécifique que x. Par ailleurs, $x \leq_{\mathcal{O}} y \Leftrightarrow x \sqcap_{\mathcal{O}} y = x$. Nous définissons ensuite l'opérateur $\sqcap_2$ tel que, pour deux descriptions X et Y de $\mathcal{D}_2$ quelconques :

$$X \sqcap_2 Y = \mathtt{max}(\leq_{\mathcal{O}}, \{x \sqcap_{\mathcal{O}} y \mid (x, y) \in X \times Y\})$$

Cette structure de patrons permet la généralisation d'EIM déclenchés par des médicaments différents qui partagent un sous-groupe pharmacologique ou chimique. Par exemple :

$$\delta(\mathrm{P1}) \sqcap_2 \delta(\mathrm{P4}) = \langle \{\{\mathrm{H02AB07}\}, \{\mathrm{N02BE01}\}\} \sqcap_2 \{\{\mathrm{H02AA03}\}\}, \emptyset \rangle$$
$$= \langle \{\mathrm{H02A}\}, \emptyset \rangle$$

Ce vecteur représente la généralisation la plus spécifique des descriptions des EIM des patients P1 et P4. Il exprime qu'au moins un médicament de la classe H02A (*corticosteroids for systemic use*, *plain*) est associé à un phénotype dans la classe ICD 580-629, et qu'aucun médicament n'est associé à la classe ICD 710-739.

3.3 Extension de la structure de patrons avec une ontologie de phénotypes

Nous définissons une troisième structure de patrons permettant d'utiliser à la fois ATC et ICD pour généraliser également les phénotypes. Ici, nous n'utilisons que les deux niveaux les plus spécifiques d'ICD, l'expérience précédente couvrant le niveau le plus général. La Table 5 illustre la représentation des données pour cette structure de patrons. Ici, les EIM sont représentés comme des vecteurs à deux dimensions $\langle D_i, P_i \rangle$, qui associe à un ensemble de médicaments D_i un ensemble de phénotypes P_i. Un patient est décrit par un ensemble d'EIM.

TABLE 5 – *Exemple de descriptions de patients pour la structure de patrons* $(G, (\mathcal{D}_3, \sqcap_3), \delta_3)$.

	Description
P1	$\{\langle\{\text{H02AB07}\},\{\text{ICD 599.8}\}\rangle, \langle\{\text{N02BE01}\},\{\text{ICD 599.9}\}\rangle\}$
P2	$\{\langle\{\text{H02AB07}\},\{\text{ICD 599.9}\}\rangle, \langle\{\text{H02AB07}\},\{\text{ICD 719.4}\}\rangle\}$
P3	$\{\langle\{\text{H02AB07}, \text{N02BE01}\},\{\text{ICD 599.9}\}\rangle, \langle\{\text{N02BE01}\},\{\text{ICD 719.4}\}\rangle\}$

On définit ainsi la structure de patrons $(G, (\mathcal{D}_3, \sqcap_3), \delta_3)$, où une description de $\mathcal{D}_3$ est un ensemble d'EIM représentés comme des vecteurs. On définit d'abord un opérateur $\sqcap_{ADE}$ sur ces vecteurs, tel que, pour deux EIM v_x et v_y :

$$v_x \sqcap_{EIM} v_y = \langle x_{ATC}, x_{ICD} \rangle \sqcap_{EIM} \langle y_{ATC}, y_{ICD} \rangle$$
$$= \begin{cases} \langle x_{ATC} \sqcap_{\mathcal{O}} y_{ATC}, x_{ICD} \sqcap_{\mathcal{O}} y_{ICD} \rangle \text{ si les deux dimensions} \\ \qquad\qquad \text{contiennent une classe non-racine,} \\ \langle \emptyset, \emptyset \rangle \text{ sinon.} \end{cases}$$

L'opérateur $\sqcap_{EIM}$ applique l'opérateur $\sqcap_{\mathcal{O}}$ sur chaque dimension du vecteur, en utilisant les ontologies ATC et ICD pour généraliser, respectivement, les médicaments et les phénotypes. Les deux dimensions du vecteur doivent contenir chacune au moins une classe non-racine de leur ontologies respectives. Si ce n'est pas le cas, on lui donne pour valeur $\langle \emptyset, \emptyset \rangle$ afin de l'ignorer : cela permet de s'assurer que l'on ne généralise que sur des EIM avec au moins un médicament et un phénotype.

On définit l'opérateur $\sqcap_3$ tel que, pour deux descriptions X et Y de $\mathcal{D}_3$ quelconques :

$$X \sqcap_3 Y = \max\left(\leq_{EIM}, \{v_x \sqcap_{EIM} v_y \mid (v_x, v_y) \in X \times Y\}\right)$$

Par rapport à $\sqcap_2$, $\sqcap_3$ ajoute une opération intermédiaire sur les EIM avec l'opérateur $\sqcap_{EIM}$, et permet d'utiliser $\sqcap_{\mathcal{O}}$ avec à la fois ATC et ICD.

3.4 Extraction de règles d'association

Ces trois expériences permettent de produire trois treillis de concepts desquels on extrait des Règles d'Association (RA). Une RA est identifiée dans le treillis entre deux concepts parents dans le treillis, de descriptions $\delta(l)$ et $\delta(r)$, avec $\delta(r)$ plus spécifique que $\delta(l)$. Une RA comporte une partie gauche $L = \delta(l)$ et une partie droite $R = \delta(r) - \delta(l)$, et est notée $L \rightarrow R$. Empiriquement, on n'extrait seulement les RA avec un support (nombre de patients vérifiant la règle) d'au moins 5 et une confiance (ratio de patients vérifiant L qui vérifient également R) d'au moins 0,75. L'extraction des RA produit un grand nombre de règles, parmi lesquelles les RA répondant à notre problème doivent être identifiées. Nous définissons donc un critère de sélection pour ces RA par la conjonction des deux conditions suivantes. *(i)* La partie droite R d'une RA contient au moins un EIM, noté $\langle D_R, P_R \rangle$, pour lequel il n'existe pas d'EIM dans la partie gauche L, noté $\langle D_L, P_L \rangle$, tel que D_L et D_R (resp. P_L et P_R) sont comparable par $\leq_{\mathcal{O}}$. Cette condition permet de s'assurer que la partie droite de la règle introduit des médicaments et phénotypes non liés à ceux de la partie droite, *i.e.*, l'association

entre les EIM n'est pas triviale. *(ii)* Puisque les patients du corpus sont diagnostiqués avec le LED, les règles ne doivent pas contenir de phénotype associé au LED (classe ICD 710 et ses descendants).

4 Résultats et discussion

Les trois expériences décrites dans cet article produisent trois treillis de concepts, desquels on extrait des Règles d'Association (RA). La Table 6 présente quelques statistiques sur les treillis et RA obtenus pour les trois expériences.

TABLE 6 – *Statistiques sur le processus d'extraction de RA pour les trois experiences, implementées en Java*.

Expérience	**1**	**2**	**3**
Taille du treillis (en millions de concepts)	1,9	2,3	2,5
RAs extraites (en millions)	5	7	9
RAs selectionnées	772	1 907	913

Nous présentons un exemple de RA obtenu dans la troisième expérience, avec un support de 10 et une confiance de 0,77 :

$$\{\{\text{C08DB01}\}, \{\text{ICD 428.0}\}\} \rightarrow \{\{\text{A02B}\}, \{\text{ICD 427.31}\}\}$$

Cette règle signifie que 77% des patients qui présentent *congestive heart failure* (ICD 428.0) après prescription de diltiazem (C08DB01), présentent également *atrial fibrillation* (ICD 427.31) après prescription d'un médicament pour l'ulcère gastroduodénal et le reflux gastro-œsophagien (A02B). Cette RA est vérifiée pour 10 patients dans notre corpus. L'ensemble complet des RA sélectionnées est disponible en ligne à l'adresse suivante : `http://www.loria.fr/~gpersone/eim-assoc/`.

Un grand nombre de RA peut être extrait de nos treillis de concepts. Nous avons automatiquement sélectionné un ensemble de ces RA en excluant les règles en dehors de la portée de notre étude. D'autres métriques pour filtrer ces RA, comme le lift, peuvent être envisagées. Nous devons maintenant classer ces RA par rapport à leur importance en termes de risques et de coûts des phénotypes présents dans leur partie droite. Une limite de cette approche est l'absence de relation temporelle entre les EIM. Nous n'avons pas considéré cet aspect, car l'ordre d'apparition d'EIM associés peut varier entre les patients. Cependant, cet ordre peut être vérifié dans les DME, puisque les concepts de patrons conservent les identifiants des patients.

Nous explorons dans cet article une approche fondée sur les structures de patrons pour extraire des EIM fréquemment associés depuis des DME. Cette approche est flexible et permet manipuler et fouiller les objets complexes que sont les EIM. De plus elle permet une généralisation sur les différents composants des EIM grâce aux ontologies médicales. La représentation des EIM pourrait être étendue en incluant d'autres composants, par exemple les cibles des médicaments annotées par des classes Gene Ontology, les dosages ou les voies d'administration.

Références

COULET, A., *ET AL.* (2013). Using pattern structures for analyzing ontology-based annotations of biomedical data. ICFCA. Springer.

GANTER, B., & KUZNETSOV, S. O. (2001). Pattern Structures and Their Projections. 9th International Conference on Conceptual Structures, (pp. 129-142).

GANTER, B., & WILLE, R. (1999). Formal Concept Analysis: Mathematical Foundations. Springer.

KUHN, M., *ET AL.* (2016). The SIDER database of drugs and side effects. Nucleic Acids Research.

LEPENDU, P., ET AL. (2013). Pharmacovigilance Using Clinical Notes. Clinical Pharmacology & Therapeutics.

LOWE, H. J., FERRIS, T. A., *ET AL.* (2009). STRIDE–An Integrated Standards-Based Translational Research Informatics Platform. AMIA Annual Symposium Proceedings, (pp. 391-395).

VASUDEVAN, A. R., & GINZLER, E. M. (2009). Established and Novel Treatments for Lupus. The Journal of Musculoskeletal Medicine.

Evaluation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie

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

[1] Equipe de Recherche en Informatique Appliquée à la Santé, INSERM 1219, Université de Bordeaux
[2] Service d'Information Médicale, Pôle de Santé Publique, CHU de Bordeaux

Résumé : Dans le domaine de la cancérologie, la réutilisation des données est confrontée à l'hétérogénéité des terminologies. Afin de répondre à cette difficulté, il est nécessaire de mettre en correspondance ces dernières. L'intégration sémantique par l'utilisation d'une troisième terminologie comme support est une approche classique pour l'alignement de deux terminologies peu structurées. Le but de notre étude était d'évaluer les capacités de la SNOMED CT à être utilisée comme support de connaissances à l'alignement de la CIMO3 et de la CIM10 en recherchant des ancrages des terminologies à aligner dans la SNOMED CT. Deux ressources différentes ont été exploitées pour identifier ces ancrages : le NCI Metathesaurus et le fichier RF2 de la SNOMED CT décrivant des correspondances entre cette dernière et des terminologies de référence. On retrouve dans les ancrages via la NCI Metathesaurus une couverture de 94,6% pour la CIMO3 et de 74,1% pour la CIM10 et via le RF2 une couverture pour les codes CIMO3 95,1% et de 84% pour la CIM10. Cette bonne couverture semble indiquer que la SNOMED CT peut constituer un bon support à l'alignement de terminologies diagnostiques en cancérologie.

Mots-clés : CIM10, CIMO3, SNOMED CT, Alignement de terminologies.

1 Introduction

L'informatisation des systèmes de santé entraine une grande production de données. L'enregistrement, la transmission et l'exploitation de l'information sanitaire ont comme problématique l'ambiguïté du langage médical. Cette ambiguïté découle du langage naturel sur lequel il repose. Le manque de consensus sur la définition d'une notion clinique, la polysémie, l'imprécision ainsi que la synonymie et la paraphrase sont autant de facteurs d'ambiguïté retrouvés dans le langage médical (Zweigenbaum, 1999). Pour pallier ces obstacles, les terminologies ont pris progressivement place dans le domaine de la santé afin de formaliser la représentation de ses connaissances. Il existe dans le domaine de la santé pratiquement autant de terminologies que de champs d'application (Joubert et al., 2009). Ces différentes terminologies, construites dans des modèles de représentations différents et pour des contextes d'utilisation différents, posent le problème de la complexité de leur interopérabilité notamment dans le cadre de la réutilisation secondaire des données de santé (Merabti et al., 2009).

La cancérologie, avec la stratégie de surveillance du cancer par les registres, est un domaine où la réutilisation des données est très importante (MacKay et Sellers, 1973). En effet, différentes sources participent à l'alimentation de ces registres (Jouhet et al., 2012). En France et plus généralement au niveau international, les registres de cancer utilisent une classification pour le codage des données qui est la Classification Internationale des Maladies pour l'Oncologie (CIMO3) (Fritz et OMS, 2008). La classification utilisée pour la production des données de prise en charge est la Classification statistique Internationale des Maladies et des problèmes de santé connexes (CIM10), notamment dans le cadre du Programme de

Médicalisation du Système d'Information (PMSI) en France (ATIH, 2015). Au niveau international, c'est l'enregistrement des causes de décès et de morbidité qui est réalisé avec la CIM10. La nécessité d'échange entre ces sources et les registres de cancer (Tognazzo et al., 2005) requiert un alignement optimal entre la CIM10 et la CIMO3.

Une des approches classiques pour l'intégration sémantique de deux terminologies est l'utilisation d'un support de connaissances qui est le plus souvent une troisième ressource terminologique. Dans cet article, nous utilisons les définitions données par Safar *et al.* (Safar et al., 2007) pour l'alignement de deux terminologies via une terminologie de support. L'idée de base est que la terminologie de support doit couvrir les domaines des terminologies à intégrer. La terminologie de support doit permettre, d'une manière indépendante, la création d'appariements entre ses propres concepts et les concepts des terminologies à aligner. Le processus de création de ces appariements est appelé la phase d'ancrage. Après la phase d'ancrage, les concepts de la terminologie de support qui participent à l'ancrage pourront être reliés en s'appuyant sur sa propre structure : c'est la phase de dérivation.

L'une des terminologies biomédicales les plus descriptives est la Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (IHTSDO, 2015a). C'est une terminologie multiaxiale née de la fusion entre deux terminologies : la Systematized Nomenclature of Medicine Reference Terminology (SNOMED RT) et la Clinical Terms Version 3 (CTV3) (Wang et al., 2002). Elle a été construite pour prendre en compte tous les aspects de la production de soins. Le mécanisme majeur de la SNOMED CT garantissant son « hyper-expressivité » est la description logique de ses concepts pré-coordonnés (IHTSDO, 2015a). Le but de notre travail est d'évaluer la possibilité de l'utiliser comme terminologie de support pour l'alignement de la CIM10 et de la CIMO3. Plus précisément, on s'intéresse ici à déterminer la couverture des connaissances de ces terminologies diagnostiques par la SNOMED CT.

2 Matériels

2.1 La SNOMED CT

Les composants de la SNOMED CT sont les concepts, les relations et les descriptions. Les concepts sont une représentation d'une notion clinique représentés par un numéro unique : le 'SCTID'. Chaque concept possède une ou plusieurs descriptions. Chaque description est un libellé humainement compréhensible du concept. Enfin, les relations, servent à créer des associations entre les différents concepts. La création d'association entre les concepts rend possible la post-coordination dans la SNOMED CT (IHTSDO, 2015a). Le format RF2 (Release Format 2) est le support actuel de mise à disposition de la SNOMED CT. Dans ce fichier, se trouvent des tables décrivant les correspondances entre la SNOMED CT et d'autres terminologies biomédicales. Notons en particulier qu'il existe des correspondances avec la CIMO3 et des correspondances avec la CIM10 (IHTSDO, 2015a). La version du RF2 utilisée dans notre étude est celle du 31 janvier 2016.

2.2 La CIM10

La CIM10 représente des entités nosologiques sous la forme d'un code alphanumérique. Les entités nosologiques sont des pathologies autonomes dans leur déterminisme, cohérentes dans leurs manifestations et organisées en fonction de leurs relations et des oppositions qu'elles partagent les unes avec les autres (Bariéty et Coury, 1963). Dans le cadre spécifique des tumeurs, le chapitre II est celui qui lui est réservé dans la CIM10. Les codes alphanumériques vont de C00 à D48 (OMS, 2009).

2.3 La CIMO3

La CIMO3 est une classification bi-axiale, topographique et morphologique. Le code topographique désigne la localisation d'origine de la tumeur et le code morphologique le type histologique. Les codes topographiques sont une adaptation des codes CIM10 consacrés aux tumeurs malignes. Les codes topographiques sont composés de quatre caractères et sont compris entre C00.0 et C80.9. Un point (.) sépare les sous-divisions des catégories à trois caractères. Les codes morphologiques comportent cinq chiffres compris entre M-8000/0 et M-9989/3. Les quatre premiers chiffres représentent le terme histologique précis. Le cinquième chiffre, derrière la barre oblique (/), est le code de comportement qui indique si la tumeur est maligne, bénigne, in situ ou de caractère malin ou bénin non assuré (Fritz et OMS, 2008).

2.4 Le NCI Metathesaurus

Le NCI Metathesaurus est un regroupement de multiples terminologies, construit suivant le modèle du Metathesaurus de l'UMLS (« NCI Metathesaurus », 2016). Le NCI Metathesaurus intègre certaines des terminologies biomédicales de l'UMLS, notamment la CIM10, ainsi que des terminologies spécifiques du cancer, comme la CIMO3. Chaque concept dans le NCI Metathesaurus possède un identifiant unique, le CUI (Concept Unique Identifier), qui regroupe les codes des terminologies sources censés représenter la même notion.

3 Méthodes

La liste exhaustive des codes CIM10 de « C00 à D48 » et des codes CIMO3 présents dans le NCI Metathesaurus (version de juin 2013) a été sélectionnée. Cette liste a été épurée des codes d'entête qui ne servent pas au codage diagnostique.

Deux ressources différentes ont été utilisées pour étudier les correspondances entre la CIM10 et la SNOMED CT et entre la CIMO3 et la SNOMED CT : le fichier RF2 fourni par la SNOMED CT et le NCI Metathesaurus. Dans les tables de correspondances proposées dans le fichier RF2, seules les correspondances actives ont été prises en compte, c'est-à-dire les correspondances décrites comme non obsolètes par la SNOMED CT. A partir du NCI Metathesaurus, nous avons sélectionné les codes SNOMED CT ainsi que les codes CIM10 et CIMO3 ayant le même CUI pour constituer parallèlement les correspondances CIM10-SNOMED CT et CIMO3-SNOMED CT. Dans les paires ainsi créées, seuls les couples impliquant des concepts SNOMED CT non obsolètes dans le RF2 ont été pris en compte.

Nous avons ensuite compté le nombre de codes CIM10 et de codes CIMO3 dans chaque source. Nous avons ainsi calculé le taux de couverture de la CIM10 et de la CIMO3 dans les correspondances trouvées. Par ailleurs, nous avons également évalué la présence des codes CIM10 et des codes CIMO3 en fonction de l'origine des correspondances (*i.e.*, via le RF2 ou via le NCI Metathesaurus).

4 Résultats

La table 1 présente la couverture des terminologies CIM10 et CIMO3 dans les correspondances retrouvées via le RF2 et via le NCI Metathesaurus.

On retrouve que via le RF2, 84 % des codes CIM10 et 95,1% des codes CIMO3 ont des correspondances avec la SNOMED CT. Via le NCI Metathesaurus, 74,1% des codes CIM10 et 94,6% des codes CIMO3 ont des correspondances avec la SNOMED CT.

Pour les codes CIM10 décrivant les hémopathies malignes, la couverture est de 70,7% via le RF2 et de 79,3% via le NCI Metathesaurus.

TABLE 1 – *Couverture des codes CIM10 et CIMO3 dans les correspondances retrouvées avec la SNOMED CT via le RF2 et via le NCI Metathesaurus. *N est le nombre de codes dans la terminologie d'origine*

			N*	RF2	NCI Meta thesaurus	RF2 ou NCI	RF2 et NCI
				Couverture	Couverture	Couverture	Couverture
Codes CIM10 (C00-D48)	Bénin (D10-D36)		180	153 *85,0%*	116 *64,4%*	163 *90,6%*	106 *58,9%*
	Hémopathie maligne (C81-C96)		92	65 *70,7%*	73 *79,3%*	87 *94,6%*	51 *55,4%*
	Imprévisible D37-D48)		86	74 *86,1%*	58 *67,4%*	80 *93,0%*	52 *60,5%*
	In situ (D00-D09)		66	56 *84,9%*	43 *65,2%*	60 *90,9%*	39 *59,1%*
	Malin primitif (C00-C75)		388	333 *85,8%*	315 *81,2%*	365 *94,1%*	283 *72,9%*
	Malin secondaire (C76-C80)		39	34 *87,2%*	25 *64,1%*	37 *94,9%*	22 *56,4%*
	Tumeurs multiples (C97)		1	1 *100,0%*	1 *100,0%*	1 *100,0%*	1 *100,0%*
	Total		**852**	**716 *84,0%***	**631 *74,1%***	**793 *93,1%***	**554 *65,0%***
Codes CIMO3	Topographiques(C_._)		330	287 *87,0%*	284 *86,1%*	305 *92,4%*	279 *84,5%*
	Morpho-logiques	Bénin (/0)	290	286 *98,6%*	284 *97,9%*	289 *99,6%*	281 *96,9%*
		Indéterminé (/1)	150	139 *92,7%*	141 *94,0%*	146 97,3%	134 *89,3%*
		In situ (/2)	30	30 *100,0%*	28 *93,3%*	30 *100,0%*	28 *93,0%*
		Malin primitif (/3)	553	545 *98,6%*	542 *98,0%*	551 *99,6%*	536 *96,9%*
		Malin secondaire (/6)	6	6 *100,0%*	6 *100,0%*	6 *100,0%*	6 *100,0%*
		Incertain si primitif ou secondaire (/9)	3	3 *100,0%*	3 *100,0%*	3 *100,0%*	3 *100,0%*
		Sous-Total	1032	1009 *97,8%*	1004 *97,3%*	1025 *99,3%*	988 *95,7%*
	Total		**1362**	**1296 *95,1%***	**1282 *94,6%***	**1330 *97,6%***	**1267 *93,0%***

Dans les correspondances pour les codes CIM10 retrouvées via le NCI Metathesaurus, en ignorant le code décrivant les tumeurs multiples, seule la couverture des tumeurs malignes primitives est au-dessus de 80%.

La figure 1 présente la répartition des codes CIM10 et des codes CIMO3 en fonction de l'origine des correspondances.

On retrouve que 65% des codes CIM10 sont mis en correspondance avec au moins un code SNOMED CT via les deux ressources utilisées. C'est le cas également de 84,5% des codes CIMO3 topographiques et 95,7% des codes CIMO3 morphologiques. Un exemple de code topographique que l'on retrouve dans les deux ressources est *C03.1 - Lower gum* en lien avec deux concepts SNOMED CT via le NCI Metathesaurus : *6912001 - Structure of lower alveolar ridge mucosa (body structure)* et *57131003 - Structure of gum of mandible (body structure)*. Via le RF2, on obtient, en plus des correspondances retrouvées via le NCI Metathesaurus, d'autres liens avec les concepts *304704007 - Entire gum of mandible (body structure)* et *368709004 - Entire lower alveolar ridge mucosa (body structure)*. Par ailleurs, on observe des correspondances fournies uniquement par une ressource.

Ainsi, le RF2 propose des correspondances pour 19% supplémentaires des codes CIM10, 2,4% des codes CIMO3 topographiques et 2% des codes CIMO3 morphologiques. Un exemple de codes CIM10 n'ayant de correspondances que via le RF2 est *C50.1 - Malignant neoplasm of central portion of breast* qui est en lien avec *93745008 - Primary malignant neoplasm of central portion of female breast (disorder)* et *448436006 - Sarcoma of central portion of female breast (disorder)*, *708921005 - Carcinoma of central portion of breast (disorder)* et *708921005 - Carcinoma of central portion of breast (disorder)*.

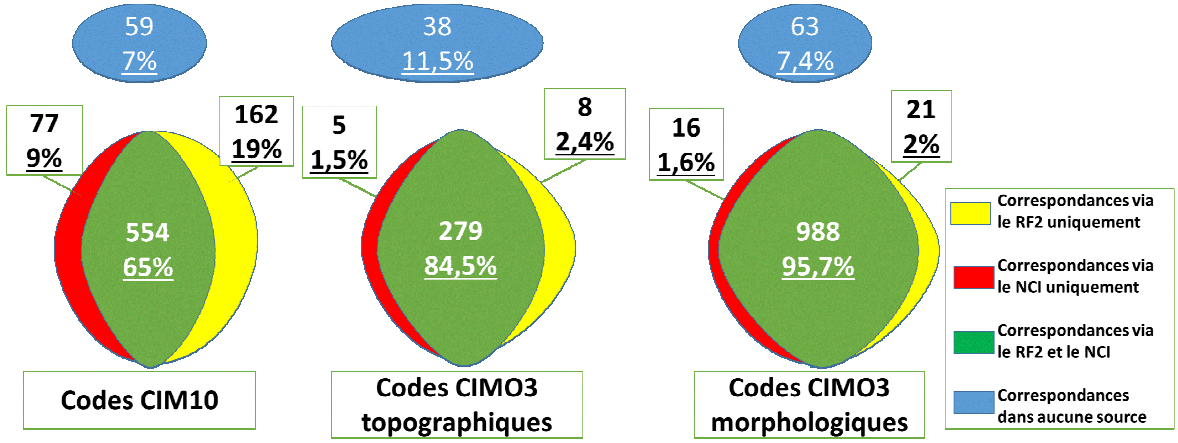


FIGURE 1 – Répartition des codes CIM10 et des codes CIMO3 en fonction de la ressource d'origine des correspondances.

Il existe également des correspondances décrites uniquement dans le NCI Metathesaurus pour 9% des codes CIM10, 1,5% des codes CIMO3 topographiques et 1,6% des codes CIMO3 morphologiques. Un exemple de code morphologique retrouvé dans le NCI Metathesaurus et pas dans le RF2 est *9260/3 - Ewing sarcoma* en lien avec *76909002 - Ewing's sarcoma (morphological abnormality)*. Un autre exemple est le code *9684/3 - Malignant lymphoma, large B-cell, diffuse, immunoblastic, NOS* en lien avec *109966003 - Diffuse non-Hodgkin's lymphoma, immunoblastic (disorder)* et *450909005 - Plasmablastic lymphoma (morphologic abnormality)*.

5 Discussion

L'utilisation d'une troisième ressource dans un processus d'alignement de deux terminologies est très pertinente notamment quand les terminologies à aligner sont faiblement structurées ou se limitent à de simples hiérarchies de classification (Safar et al., 2007). La CIM10 et la CIMO3 étant des classifications, il semblait judicieux d'exploiter une terminologie de support capable d'intégrer les notions cliniques qui y sont représentées. De plus, les codes CIM10, représentant des pathologies dans une structure mono-axiale, ne peuvent en aucun cas être directement mis en correspondance avec des codes CIMO3 puisque ces derniers représentent des sites anatomiques et des lésions histologiques tumorales dans une structure indépendante et donc bi-axiale. La SNOMED CT, grâce à la description logique des concepts pré-coordonnés, permet de trouver des relations entre des concepts de pathologie tumorale, leur site et leur lésion histologique (Bodenreider, 2015). Cette propriété de la SNOMED CT rend possible les dérivations en son sein. L'existence d'ancrages entre la CIM10 et la SNOMED CT ainsi qu'entre la CIMO3 et la SNOMED CT indique que la SNOMED CT pourrait bien jouer le rôle de terminologie de support en vue de l'alignement de la CIM10 et de la CIMO3. Il faut noter qu'une alternative à la SNOMED CT comme support est le NCI Thesaurus, comme cela a été fait dans des études précédentes (Brechat et al., 2014 ; Burgun et Bodenreider, 2007). Une perspective intéressante serait la comparaison de l'alignement obtenu entre la CIM10 et la CIMO3 via ses deux ressources.

Via le RF2 et le NCI Metathesaurus, nous avons retrouvé un taux de couverture générale assez élevé dans la participation aux ancrages pour la CIM10 et la CIMO3. Des correspondances existent pour la majorité des codes CIM10 et CIMO3, quelle que soit la technique utilisée. La couverture la plus importante (97%) est celle des codes CIMO3 morphologiques, ce qui s'explique par le fait qu'ils ont été utilisés comme support pour la représentation des lésions morphologiques tumorales dans la SNOMED CT (IHTSDO, 2016).

En plus de cette spécificité pour les codes CIMO3 morphologiques, la couverture générale élevée pourrait être consécutive à l'étendue du spectre de spécialités cliniques couvertes par la SNOMED CT et à la granularité des descriptions de chaque spécialité clinique. Pour notre étude, la SNOMED CT semble répondre ainsi aux critères d'une terminologie de support en contenant les connaissances et les descriptions de la CIM10 et de la CIMO3.

On note spécifiquement une faible couverture, quelle que soit la source utilisée, des codes décrivant les hémopathies malignes. Cela pourrait témoigner de la complexité et de la forte évolutivité des connaissances de ce sous-domaine. Pour la suite de notre étude, cela pourrait entrainer un alignement moins satisfaisant dans ce sous-domaine. Pourtant, les hémopathies représentent une part importante des pathologies cancéreuses (Monnereau et al., 2013).

Le nombre de correspondances trouvées via le NCI Metathesaurus est plus faible que via le RF2 mais certaines d'entre elles n'existent pas dans le RF2. Le code CIMO3 *9260/3 - Ewing sarcoma* en est une illustration. Cela peut s'expliquer par le fait que, dans le RF2, les correspondances sont orientées de la SNOMED CT vers les terminologies cibles (IHTSDO, 2015b), l'essentiel étant de couvrir au mieux les codes SNOMED CT. Cette caractéristique des correspondances proposées par le RF2 pose ainsi une difficulté pour notre objectif.

Créer des correspondances sémantiques entre concepts de deux terminologies revient à les intégrer virtuellement (Klein, 2001). La stratégie d'alignement utilisée dans le NCI Metathesaurus, basée sur une approche morphosyntaxique (Schuyler et al., 1993), peut entrainer des inconsistances dans les alignements proposés. C'est le cas pour le code CIMO3 morphologique *9684/3 - Malignant lymphoma, large B-cell, diffuse, immunoblastic, NOS* qui est en lien avec le code SNOMED CT de maladie *109966003 - Diffuse non-Hodgkin's lymphoma, immunoblastic (disorder)*. En effet, un code CIMO3 morphologique représente une lésion histologique telle que le décrirait un anatomopathologiste et le code SNOMED CT décrit la maladie ayant cette lésion histologique comme caractéristique. Ces deux concepts sont donc liés pas équivalents. Pour éviter ce type d'inconsistances, nos prévoyons d'exploiter la hiérarchie de la SNOMED CT. En particulier, les codes CIM10 devraient être associés à des concepts SNOMED CT de type *disorder*, tandis que les codes CIMO3 morphologiques devraient être associés à des concepts SNOMED CT de type *morphologic abnormality* et les codes CIMO3 topographiques à des concepts SNOMED CT de type *body structure*.

Pour les codes ayant des correspondances dans les deux ressources exploitées, des correspondances communes aux deux sources ont été obtenus. Chaque stratégie de création de correspondances dans les différentes sources présente des avantages et des difficultés pour nos objectifs. Il est donc judicieux pour nous d'utiliser les deux sources de correspondances afin de pallier les défauts des différentes stratégies de correspondances et ainsi d'aligner d'une manière optimale le maximum de concepts. En revanche, il nous sera nécessaire de prévoir la gestion des incohérences susceptibles d'apparaitre entre les deux sources de correspondances.

La prochaine étape de notre travail consistera donc à prendre en compte le nombre de concepts SNOMED CT mis en correspondance avec chaque code CIM10 ou code CIMO3, de filtrer les correspondances impliquant des inconsistances et enfin de procéder à des dérivations en utilisant les descriptions logiques des concepts de la SNOMED CT.

6 Conclusion

Cette étude a permis de mettre en évidence la couverture suffisante de la SNOMED CT pour servir de support à l'alignement de deux terminologies diagnostiques dans le domaine de la cancérologie : la CIM10 et la CIMO3. Il a été précédemment démontré que la SNOMED CT, grâce à la description logique de ses concepts pré-coordonnés, permet d'améliorer la cohérence et la consistance dans les alignements de terminologies quand elle joue le rôle de terminologie de support (Brown et al., 2007). La prochaine étape de notre étude visera donc à analyser la qualité des correspondances proposées par chaque source et à utiliser la description logique dans la SNOMED CT pour procéder à la phase de dérivation afin d'obtenir un alignement optimal entre la CIM10 et la CIMO3.

Références

AGENCE TECHNIQUE DE L'INFORMATION SUR L'HOSPITALISATION. (2015). Classification statistique internationale des maladies et des problèmes de santé connexes CIM-10 FR à usage PMSI. MINISTÈRE DES AFFAIRES SOCIALES ET DE LA SANTÉ. http://www.atih.sante.fr/sites/default/files/public/content/2665/cim10_2015_final_0.pdf

BARIETY, M., & COURY, C. (1963). *Histoire de la médecine* (p. 625). Paris: Fayard.

BODENREIDER, O. (2015, mai). *Oncology in SNOMED CT*. Bethesda, Maryland. https://mor.nlm.nih.gov/pubs/pres/20150513-CancerBigData.pdf

BRECHAT, B., MOUGIN, F., THIESSARD, F., & JOUHET, V. (2014). Mapping de terminologies diagnostiques en cancérologie par l'intermédiaire du NCI Metathesaurus. *Actes des 15es Journées francophones d'informatique médicale (JFIM 2014)*, 34-43.

BROWN, S. H., HUSSER, C. S., WAHNER-ROEDLER, D., BAILEY, S., NUGENT, L., Porter, K., ELKIN, P. L. (2007). Using SNOMED CT as a reference terminology to cross map two highly pre-coordinated classification systems. *Studies in Health Technology and Informatics*, *129*(Pt 1), 636-639.

BURGUN, A., & BODENREIDER, O. (2007). Issues in integrating epidemiology and research information in oncology: experience with ICD-O3 and the NCI Thesaurus. *AMIA Annual Symposium Proceedings*, 85-89.

FRITZ, A., & ORGANISATION MONDIALE DE LA SANTE. (2008). *Classification internationale des maladies pour l'oncologie*. Genève: Organisation mondiale de la santé.

IHTSDO. (2015a). SNOMED CT® Technical implementation Guide January 2015 International Release (GB English). http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_TechnicalImplementationGuide_Current-en-GB_INT_20150131.pdf

IHTSDO. (2015b, janv 31). Mapping SNOMED CT to ICD-10 Technical Specifications. https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm_tech_spec_20130601.pdf

IHTSDO. (2016). SNOMED CT® Editorial Guide January 2016 International Release (US English). https://confluence.ihtsdotools.org/download/attachments/5505533/IHTSDO Editorial Guide 20160131.pdf

JOUBERT, M., DAHAMNA, B., DELAHOUSSE, J., FIESCHI, M., & DARMONI, S. J. (2009). SMTS®: un serveur multiterminologies en santé. *Risques, Technologies de l'Information pour les Pratiques Médicales*, 47–56.

JOUHET, V., DEFOSSEZ, G., BURGUN, A., LE BEUX, P., LEVILLAIN, P., INGRAND, P., & CLAVEAU, V. (2012). Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of Information in Medicine*, *51*(3), 242-251.

KLEIN, M. (2001). Combining and relating ontologies: an analysis of problems and solutions. *IJCAI-2001 Workshop on ontologies and information sharing* (p. 53–62).

MACKAY, E. N., & SELLERS, A. H. (1973). The Ontario cancer incidence survey, 1964-1966: a new approach to cancer data acquisition. *Canadian Medical Association Journal*, *109*(6), 489.

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*, 79-88.

MONNEREAU, A., REMONTET, L., MAYNADIE, M., BINDER-FOUCARD, F., BELOT, A., TROUSSARD, X., & BOSSARD, N. (2013). *Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012: étude à partir des registres des cancers du réseau Francim*. Saint-Maurice: Institut de veille sanitaire.

NCI METATHESAURUS. (2016). https://ncimeta.nci.nih.gov/ncimbrowser/

ORGANISATION MONDIALE DE LA SANTE. (2009). *Classification statistique internationale des maladies et des problèmes de santé connexes, Dixième version*. https://www.cihi.ca/fr/icd_volume_one_2012_fr.pdf

SAFAR, B., REYNAUD, C., & CALVIER, F. (2007). Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire. *Actes des 1ères Journées Francophones sur les Ontologies (JFO 2007)*, 21–35.

SCHUYLER, P. L., HOLE, W. T., TUTTLE, M. S., & SHERERTZ, D. D. (1993). The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, *81*(2), 217.

TOGNAZZO, S., ANDOLFO, A., BOVO, E., FIORE, A. R., GRECO, A., GUZZINATI, S., ZAMBON, P. (2005). Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry. Quality control of cancer cases automatically registered. *European Journal of Public Health*, *15*(6), 657-664.

WANG, A. Y., SABLE, J. H., & SPACKMAN, K. A. (2002). The SNOMED clinical terms development process: refinement and analysis of content. *AMIA Annual Symposium Proceedings*, 845-849.

ZWEIGENBAUM, P. (1999). Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, 2-3, 27-47.