PROPOSITION DE SUJET DE THESE

SIGLE ET NOM DU LABORATOIRE : LIMICS, UMRS 1142
NOM DE L'EQUIPE : LABORATOIRE D'INFORMATIQUE MEDICALE ET D'INGENIERIE DES CONNAISSANCES EN ESANTE
DIRECTEUR DE THESE : XAVIER TANNIER
CO-ENCADRANT : CHRISTEL DANIEL (AP-HP WIND, LIMICS)
LIEU DE LA THESE : 15, RUE DE L'ECOLE DE MEDECINE, 75006 PARIS

TITRE DE LA THESE : IDENTIFICATION DE PHENOTYPES A GRANDE ECHELLE PAR APPRENTISSAGE SEMI-SUPERVISE

PREREQUIS, FORMATION : FORMATION D'INFORMATIQUE, CONNAISSANCES AVANCEES EN APPRENTISSAGE AUTOMATIQUE, NOTIONS EN TRAITEMENT AUTOMATIQUE DES LANGUES.
CONTACT POUR CE SUJET : XAVIER TANNIER, CHRISTEL DANIEL
EMAIL : XAVIER.TANNIER@SORBONNE-UNIVERSITE.FR, CHRISTEL.DANIEL@APHP.FR
TELEPHONE : 01 44 27 91 13

PRESENTATION DU SUJET

Mots-clés : Identification de phénotype, sélection de cohortes, apprentissage semi-supervisé, réseaux de neurones, traitement automatique des langues.

Contexte

L'analyse des dossiers patients informatisés peut conduire à de nombreux bénéfices pour le soin et la recherche clinique (Longhurst *et al.*, 2014 ; Pathak *et al.*, 2013 ; Shah, 2013 ; Shivade *et al.*, 2013). Parmi ces bénéfices, l'aide à la constitution de cohortes de patients partageant des caractéristiques cliniques ou biologiques communes (phénotypage) peut permettre de faciliter et d'accélérer le travail des chercheurs (Gottesman *et al.*, 2013 ; Wei and Denny, 2015), voire même de conduire à de nouvelles découvertes cliniques (Denny *et al.*, 2013 ; Carroll *et al.*, 2015 ; Ritchie *et al.*, 2014 ; Lin *et al.*, 2015).
De nombreux travaux de formalisation et de classification des critères d'éligibilité à des essais cliniques ont été réalisés, dans le but d'obtenir des représentations formelles de ces critères et de faciliter l'identification des patients éligibles dans le cadre d'études de faisabilité d'essais cliniques ou d'aide informatisée au recrutement (Weng *et al.*, 2010; Luo *et al.*, 2011; Shivade *et al.*, 2014; Daniel *et al.*, 2016). La sélection de cohortes se fait au quotidien par des requêtes structurées et complexes sur des outils spécifiques (Doods *et al.*, 2014; Soto-Rey *et al.*, 2015; Girardeau *et al.*, 2017; Jiang *et al.*, 2016). La base de données PheKB (Phenotype KnowledgeBase) contient des définitions de phénotypes à base de règles intégrant des concepts codés (en utilisant SNOMED, LOINC, ICD10, etc.) ainsi que les performances des requêtes correspondantes implémentées au sein des entrepôts de données de santé des institutions partenaires (Kirby *et al.*, 2016). Par ailleurs, de nombreuses approches statistiques de traitement automatique des langues ont été proposées récemment pour exploiter les textes des dossiers patient afin de permettre l'identification de phénotypes. La plupart de ces méthodes tombent dans la catégorie de l'apprentissage totalement supervisé, c'est-à-dire qu'elles nécessitent la création d'un jeu de données déjà annoté (par des experts humains) pour permettre l'entraînement d'un modèle statistique, qui ensuite pourra être appliqué sur des données nouvelles. Ces approches demandent donc un investissement de départ considérable, et sont difficiles à généraliser car les annotations manuelles sont spécifiques à un cas particulier ; c'est la raison pour laquelle elles n'ont été appliquées que sur un nombre de phénotypes relativement limité.
Plus récemment, des approches dites "semi-supervisées" ont été proposées, permettant de s'affranchir partiellement de l'étape d'annotation manuelle, la remplaçant tantôt par un mécanisme d'amorçage (définition d'exemples ou de termes très discriminants pour un phénotype considéré, permettant de servir d'amorces à un processus itératif de sélection de documents), tantôt par un apprentissage dit "actif", dans lequel les dossiers proposés à l'expert sont choisis automatiquement de manière à minimiser le nombre de patients à explorer avant d'obtenir un jeu de données d'entraînement de bonne qualité (Halpern *et al.*, 2016; Agarwal *et al.*, 2016; Beaulieu-Jones *et al.*, 2016). Ces progrès récents ont été exclusivement effectués sur des documents de langue anglaise. Or, cette langue est dotée d'outils de traitement et de ressources terminologiques bien supérieures aux autres, et les approches ne sont pas directement transposables au français par exemple. En français, les travaux sont nombreux sur les textes du domaine général, beaucoup moins sur le domaine biomédical (Névéol *et al.*, 2018).

Par ailleurs, l'identification de phénotypes à grande échelle ("high-throughput phenotyping", "next generation phenotyping") nécessite des méthodes permettant l'exploitation conjointe des données structurées et non structurées

telles que les documents textuels mais également les images, les signaux temps réel. Utiliser l'ensemble des informations et des connaissances disponibles nous semble être une condition nécessaire pour s'approcher des performances d'un expert humain pour la sélection de patients. L'exploitation de données aussi hétérogènes est un défi, mais des progrès récents dans le domaine de la représentation d'information, notamment grâce aux réseaux de neurones, montrent que l'on peut représenter de façon jointe tous les types de structure dans un même espace, permettant la mise en œuvre d'algorithmes sur un seul mode de représentation issu de multiples sources. Ce type d'approches a été appliqué par exemple à la représentation jointe d'images et de texte (Socher *et al.*, 2014), de bases de connaissances et de textes (Sun *et al.*, 2015), et très récemment de dossiers patients avec données structurées et texte (Miotto *et al.*, 2016).

Objectifs et méthodes

Notre objectif est donc de traiter trois questions distinctes liées à l'automatisation de l'identification de phénotypes et de la constitution de cohortes utilisant les nouvelles technologies d'intelligence artificielle :

1. L'adaptation d'algorithmes existants à la langue française, moins dotée et moins étudiée
2. L'utilisation conjointe du texte et des données structurées
3. La réduction de la supervision nécessaire par des approches semi-supervisées et la mise au point d'une méthode généralisable à l'ensemble ou à la grande majorité des définitions de phénotypes

Dans un souci de gestion des risques liés à l'ambition du projet, mais aussi d'identification précise des problèmes à considérer, ces trois questions seront traitées l'une après l'autre. Dans cette optique, nous proposons trois cas d'étude qui présentent tous un intérêt important pour la santé publique et permettront au doctorant d'avancer sur des bases solides et de valoriser régulièrement son travail.

Le **premier cas d'étude** est l'identification de patientes présentant des lésions suspectes ou prouvées de cancer du sein par analyse automatique de comptes-rendus de radiologie ou d'anatomo-pathologie mammaires. Dans le cadre du dépistage du cancer du sein, les radiologues utilisent la classification Birads (Breast Imaging Reporting And Data System) pour classer en 6 niveaux la morphologie du sein (de "normal" à "cancer prouvé") et définir en fonction la conduite à tenir - surveillance ou biospie. Les radiologues évaluent régulièrement leur performance diagnostique en comparant les diagnostics radiologiques et anatomo-pathologiques (gold standard). Par ailleurs, des recherches en analyse d'image ont pour but de réaliser une classification diagnostique automatique des lésions mammographiques (Mehdy *et al.*, 2017). Il s'agira de réaliser une catégorisation automatique i) de comptes-rendus d'imagerie mammaire de dépistage selon la classification Birads et ii) de comptes-rendus d'anatomie pathologique selon le diagnostic morphologique. Les outils développés aideront les radiologues d'une part à évaluer de manière continue leur performance diagnostique et d'autre part à constituer des cohortes de patientes dont les mammographies ou tomosynthèses constitueront des jeux d'apprentissage pour le développement d'algorithmes de dépistage automatique. L'approche adoptée sera totalement supervisée et uniquement basée sur le texte des comptes-rendus. Seul le **verrou 1** cité ci-dessus sera donc appréhendé, à savoir la prise en compte des spécificités de la langue médicale française pour la classification de documents.
Évaluation : Cette question a été étudiée récemment pour la langue anglaise (Castro *et al.*, 2017), ce qui permettra une comparaison indirecte, même si les corpus utilisés ne sont pas disponibles. Nous commencerons donc par adapter les approches décrites dans Castro *et al.* (2017), à savoir des méthodes de classification supervisée traditionnelle (CRF, SVM, Naive Bayes), avant de les comparer à des approches plus récentes à base de réseaux de neurones, que nous avons déjà appliquées à des comptes-rendus médicaux avec succès (Tourille *et al.*, 2017a,b) tout en tenant compte de la structure et de la composition particulière des textes en sélectionnant les portions de texte d'intérêt.

Le **second cas d'étude** est l'identification de patients atteints de diabète de type 2. Le diabète de type 2 est l'un des 30 phénotypes de la base PheKB, décrite ci-dessus. Nous nous proposons d'explorer l'opportunité de tirer parti des données structurées (**verrou 2**), dans un cadre contraint et précis. Nous utiliserons et adapterons à la pratique française la définition du diabète de type 2 formalisée par PheKB associant des codes diagnostiques CIM-10, des codes LOINC de résultats d'analyses biologiques et des codes ATC de médicaments afin de composer un ensemble d'apprentissage d'une fiabilité relative mais correcte. Un algorithme d'identification de diabétiques de type 2 exploitant à la fois les données structurées - données démographiques des patients, diagnostics et actes PMSI, résultats de laboratoire et de prescription médicamenteuse - et non structurées - comptes rendus médicaux - de dossiers patients informatisés sera développé. Pour cela, notre objectif sera de représenter ces données hétérogènes dans un même espace vectoriel, de façon jointe, en nous inspirant des travaux réalisés dans le domaine général (par exemple, Socher *et al.*, 2014 ; Sun *et al.* 2015), puis d'appliquer sur ces représentations des algorithmes à base de réseaux de neurones pour obtenir un modèle de classification. Notons que les vecteurs issus des données structurées et du texte sont à la fois des entrées et des paramètres du modèle, puisque la phase d'apprentissage permet de modifier les représentations elles-mêmes (par rétropropagation du gradient), et donc de "sélectionner" les données pertinentes pour chaque problème spécifique.
Évaluation : Les performances de l'algorithme d'identification de diabétiques de type 2 seront comparées à celles des systèmes similaires publiés (Pathak *et al.*, 2012 ; Anderson *et al.*, 2016 ; Agarwal *et al., 2016* ; Kagawa *et al.*, 2017 ; Zheng *et*

al., 2017). Une évaluation comparative selon la langue de travail sera réalisée en appliquant l'algorithme d'identification de diabétiques de type 2 à la base de données américaines MIMIC (Saeed *et al.*, 2002).

Enfin, l'objectif principal et final du projet de thèse est de concevoir une méthode et **un système pour la sélection semi-supervisée de cohortes (verrou 3)**. En particulier, nous souhaitons développer une approche généralisée, applicable à l'ensemble ou à la grande majorité des critères. Dans cette optique, nous pensons qu'il est illusoire dans un futur proche de vouloir se passer totalement de l'intervention humaine, mais qu'il est possible de la réduire considérablement et ainsi que faire gagner du temps aux chercheurs. La méthode que nous proposons se situe ainsi à mi-chemin entre les approches par amorçage et les approches d'apprentissage actif. D'une part, de manière similaire à l'approche par amorçage de Halpern *et al.* (2016), il s'agira de générer de manière semi-automatique sur la base d'une liste de mots-clés spécifiques aux phénotypes d'intérêt des données d'apprentissage labellisées bruitées (silver standard) puis de développer l'extracteur de phénotypes. D'autre part, pour parvenir à de bonnes performances générales sur tous les types de critères, on peut imaginer qu'une dose d'apprentissage actif permettant, à chaque itération, de faire sélectionner par un expert humain les expressions ou les zones de données les plus discriminantes, conduira à des résultats intéressants avec une intervention humaine limitée (valider des éléments isolés étant beaucoup moins chronophage que lire et annoter des comptes-rendus entiers).
Évaluation : Le système de sélection semi-supervisé de cohortes sera testé dans le cas du diabète de type 2, de l'infarctus du myocarde et du cancer du sein et ses performances seront comparées à celle obtenues par Agarwal *et al*. (2016). Il sera surtout évalué grâce à plusieurs dizaines de cohortes déjà mises en place au sein de l'AP-HP, pour lesquelles les critères et les patients sélectionnés sont disponibles.

Accès aux données de l'EDS de l'AP-HP
L'Entrepôt de Données de Santé (EDS) de l'AP-HP intègre des données administratives et médicales de plus de 8 millions de patients hospitalisés. Une demande d'accès aux données de l'EDS auprès du Comité Scientifique et Éthique (CSE) permettra de disposer d'un environnement de travail intégrant les données nécessaires à la recherche (données démographiques des patients, diagnostics et actes PMSI, résultats de laboratoire et de prescription médicamenteuse, comptes rendus médicaux et en particulier de radiologie et d'anatomie pathologique).

Résultats attendus
La levée des verrous cités ci-dessus conduira à trois types de résultats :

- Du point de vue de la recherche en informatique et plus spécifiquement en traitement automatique des langues et en extraction d'information, l'application généraliste d'algorithmes semi-supervisés sur des données hétérogènes en est à ses débuts, y compris dans le domaine général, et totalement inédite pour la langue française. Nos résultats constitueront des avancées dans le domaine.
- Du point de vue médical, une assistance automatisée pour la collecte et l'annotation de données doit permettre d'améliorer et d'accélérer les études observationnelles sur données mais également les études interventionnelles en facilitant les études de faisabilité d'essais cliniques et les campagnes de recrutement.
- Dans le contexte de l'entrepôt de données de santé (EDS) de l'APHP, où nos travaux seront évalués, l'amélioration des performances des outils de constitution de cohortes est nécessaire pour que l'EDS de l'AP-HP offre les conditions d'émergence d'aide décisionnelle à partir de données exploitant les nouvelles technologies de l'intelligence artificielle.

Références
Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc JAMIA. nov 2016;23(6):1166-73.
Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. déc 2016;64:168-78.
Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform. 2017;69:177-87.
Daniel C, Ouagne D, Sadou E, Forsberg K, Gilchrist MM, Zapletal E, et al. Cross border semantic interoperability for clinical research: the EHR4CR semantic resources and services. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2016;2016:51-9.
Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. déc 2013;31(12):1102-10.
Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F, et al. Piloting the EHR4CR feasibility platform across Europe. Methods Inf Med. 2014;53(4):264-8.
Girardeau Y, Doods J, Zapletal E, Chatellier G, Daniel C, Burgun A, et al. Leveraging the EHR4CR platform to support patient inclusion in academic studies: challenges and lessons learned. BMC Med Res Methodol. 28 févr 2017;17(1):36.
Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med Off J Am Coll Med Genet. oct 2013;15(10):761-71.
Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc JAMIA. juill 2016;23(4):731-40.
Jiang G, Kiefer RC, Rasmussen LV, Solbrig HR, Mo H, Pacheco JA, et al. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. J Biomed Inform. août 2016;62:232-42.
Kagawa R, Kawazoe Y, Ida Y, Shinohara E, Tanaka K, Imai T, et al. Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach. J Diabetes Sci Technol. juill 2017;11(4):791-9.

Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc JAMIA. nov 2016;23(6):1046-52.
Longhurst CA, Harrington RA, Shah NH. A « green button » for using aggregate patient data at the point of care. Health Aff Proj Hope. juill 2014;33(7):1229-35.
Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. J Biomed Inform. déc 2011;44(6):927-35.
Luo L, Li L, Hu J, Wang X, Hou B, Zhang T, et al. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. BMC Med Inform Decis Mak.;16:114.
Mehdy MM, Ng PY, Shair EF, Saleh NIM, Gomes C. Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer. Comput Math Methods Med. 2017;2017:2610628.
Miotto, R.; Li, L.; Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records Scientific Reports, 2016, 6
Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semantics.2018 Mar 30;9(1):12.
Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes. AMIA Annu Symp Proc AMIA Symp. 2012;2012:699-708.
Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc JAMIA. déc 2013;20(e2):e206-211.
Ritchie MD, Verma SS, Hall MA, Goodloe RJ, Berg RL, Carrell DS, et al. Electronic medical records and genomics (eMERGE) network exploration in cataract: several new potential susceptibility loci. Mol Vis. 2014;20:1281-95.
Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol. 2002;29:641-4.
Shah NH. Mining the ultimate phenome repository. Nat Biotechnol. déc 2013;31(12):1095-7.
Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc JAMIA. avr 2014;21(2):221-30.
Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D. & Ng, A. Y. Grounded Compositional Semantics for Finding and Describing Images with Sentences Transactions of the Association for Computational Linguistics, 2014
Soto-Rey I, Trinczek B, Girardeau Y, Zapletal E, Ammour N, Doods J, et al. Efficiency and effectiveness evaluation of an automated multi-country patient count cohort system. BMC Med Res Methodol. 1 mai 2015;15:44.
Sun, Y.; Lin, L.; Tang, D.; Yang, N.; Ji, Z. & Wang, X. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), 2015, 1333-1339
Tourille J, Ferret O, Tannier X, Névéol A. LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives. in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017). Vancouver, Canada, August 2017.
Tourille J, Ferret O, Tannier X, Névéol A. Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers. in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, August 2017.
Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc JAMIA. avr 2016;23(e1):e20-27.
Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. juin 2010;43(3):451-67.
Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inf. janv 2017;97:120-7.