Portabilité des algorithmes de phénotypage: le cas de la polyarthrite rhumatoïde dans le dossier patient informatisé en français

Thibaut Fabacher, Erik André Sauleau, Noémie Leclerc Du Sablon, Hugo Bergier, Jacques-eric Gottenberg, Adrien Coulet, Aurélie Névéol

Portabilité des algorithmes de phénotypage: le cas de la polyarthrite rhumatoïde dans le dossier patient informatisé en français

T. Fabacher[1,2,3,4], **E.-A. Sauleau** [1,2], **N. Leclerc du Sablon** [1], **H. Bergier**[1], **J.-E. Gottenberg**[1], **A. Coulet,** [3,4,†], **A. Névéol**[5,†]

[1]**University hospital, Strasbourg, France;** [2]**Icube Laboratory, Strasbourg, France**

[3]**Inria Paris, Paris, France**

[4]**Centre de Recherche des Cordeliers, Inserm, Paris, France**

[5]**Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France**

[†]**Co-last-authors**

Introduction Les dossiers patients informatisés (DPI) permettent une réutilisation secondaire des données des hôpitaux, et en particulier la conception et la réalisation d'études cliniques rétrospectives. L'une des premières étapes de ces études cliniques est la définition d'une cohorte de patients qui partagent une caractéristique ou une pathologie particulière. Cette tâche est généralement appelée *phénotypage électronique* (ou phénotypage) et se révèle souvent plus complexe qu'une simple requête par mot clef [1,2].

Une des difficultés rencontrées pour la définition des cohortes vient de la nature complexe des DPI, qui contiennent des données hétérogènes, incomplètes, structurées et non structurées, sur des périodes de temps de longueur variable et discontinues. Ainsi, la recherche d'un trait phénotypique peut nécessiter la considération à la fois des champs structurés, des textes non structurés et des marqueurs temporels. La composante temporelle est importante à prendre en compte, elle permet de définir un niveau de granularité supplémentaire des traits phénotypique, au niveau par exemple d'un patient, ou d'un séjour. Une autre difficulté vient du fait que les algorithmes de phénotypage peuvent ne pas être bien transférables d'un contexte clinique à un autre. En effet, les variations dans la collecte des données, la pratique clinique, le codage des actes médicaux, les langues font qu'un algorithme de phénotypage développé pour un lieu peut nécessiter une adaptation importante pour être transféré dans un nouveau cadre clinique.

Le phénotypage soulève plusieurs défis en matière de traitement automatique du langage (TAL). Il nécessite l'extraction d'éléments d'information rares à partir de grandes quantités de textes, ainsi que la définitions de phénotypes à partir de ces éléments d'information . L'extraction d'information à partir de texte libre demeure une tâche complexe et cette complexité est majorée parle caractère très spécifique des textes cliniques. En effet, le langage clinique se différencie du langage présent dans les articles de presse ou les romans notamment par la présence de nombreuses négations, nombreuses abréviations, le vocabulaire scientifique utilisé, des phrases ne suivant pas une construction grammaticale classique et l'importante présence de liste d'items. De plus l'aspect très sensible des données utilisées rend leur partage difficile. Il est donc nécessaire de développer des processus transférables et reproductibles[3].

Dans notre travail, nous étudions particulièrement la portabilité d'algorithmes de phénotypage de la polyarthrite rhumatoïde (PR), une pathologie auto-immune chronique qui affecte principalement les articulations. Cette pathologie est majoritairement suivie à l'hôpital lors de consultations spécialisées pour les patients complexes. Nous nous intéressons à la PR parce qu'il s'agit d'une pathologie fréquente, qu'elle est actuellement associée à de nombreuses questions cliniques qui pourraient bénéficier d'outil d'aide à la décision clinique s'appuyant sur le TAL (par exemple prédire le pronostic du patient ou les meilleures options de traitement). De plus, plusieurs algorithmes de phénotypage pour la PR ont été décrits dans la littérature et la question se pose sur leur capacité à être transférable [4,5].

Plus précisément, notre objectif est d'évaluer l'adaptabilité des algorithmes de phénotypage de la PR à un nouvel hôpital, tant au niveau d'un patient que d'un séjour à l'hôpital (hospitalisation ou consultation médicale).

Méthodes Deux algorithmes sont adaptés au contexte du CHU de Strasbourg et évalués à l'aide d'un nouveau corpus de référence de la PR annoté manuellement sur une période allant de 2015 à 2020. Pour l'évaluation des performances des modèles, une cohorte de validation annotée manuellement a été réalisée. Un ensemble de 140 patients a été annoté au niveau de la séjour. Pour chacun des séjours (~ 1000) de chacun des patients, deux médecins ont annoté si la séjour était en rapport direct ou non avec une PR. Si au moins un des séjours entre 2015 et 2020 est en rapport avec une PR,

le patient est considéré comme étant PR+.

Les deux algorithmes sont comparés à un algorithme naïf de phénotypage (appelé Baseline) qui s'appuye seulement sur des mots clés et sur les codes CIM-10 du PMSI. PMSI (Programme de Médicalisation des Systèmes d'Information). Dans le cadre du PMSI, des codes CIM-10 sont attribués à chaque séjour hospitalier pour décrire les pathologies, ce qui donne une information « gros grain »sur la raison principale et les raisons secondaires du séjour d'un patient. Cet algorithme naif est également enrichis par la détection simple du contexte des mots clefs recherchés.La considération du contexte permet de filtrer les variantes hypothétiques, en rapport avec un autre membre de la famille ou négativés des termes recherchés.

Le premier algorithme, appelé Carroll suit une approche supervisée[4] dont le modèle est pré entraîné sur les données d'hôpitaux américains. L'algorithme supervisé consiste à l'utilisation d'une régression logistique qui à partir d'information extraite du DPI donne une probabilité pour chaque patient d'être atteint de PR. Les données comprennent des données de biologies et des codes CIM-9 renseignés de façon structurée et des données extraites des textes cliniques de chaque patient. Pour l'extraction d'information des données à partir de texte libre, un ensemble d'expression régulière à été développé dans les hôpitaux américains. Afin d'adapter l'algorithme au contexte local, une traduction codifie CIM-9 vers code CIM-10 et une traduction des expressions régulière a été nécessaire. Cette traduction des expressions régulières s'est faite en deux temps. Le premier consistait en une traduction littérale des expressions régulières à l'aide de terminologie bilingue (UMLS). Cette étape a été suivie d'une évaluation des expressions sur les données locales et une adaptation de ces expressions par rapport aux données cliniques. Le second algorithme, appelé PheVis, suit une approche semi-supervisée.[5] Il prend en entrée des entités nommées extraites du dpi et des codes CIM-10. Il se base sur une approche semi-supervisée où un premier score simple (à base de règle) est calculé pour l'ensemble des patients et les patients ayant une très haute ou très faible probabilité d'être PR+ sont utilisés pour entraîner un modèle de classification utilisant l'ensemble des données d'entrée. L'extraction des entités des textes cliniques a été faite par une approche à base de dictionnaire[6]. Nous proposons également une amélioration dans la définition du silver standard.

Résultats Les algorithmes adaptés offrent des performances comparables entre eux. Pour le phénotypage au niveau du patient sur le nouveau corpus les résultats sont prometteurs (F1 0.71 à 0.79). Mais les performances sont plus faibles pour le phénotypage au niveau du séjour (F1 0.54 à 0.57).

Methods	Prec.	NPV	Spe.	Rec.	bal Acc.	Acc.	F1*	AUC*
CIM-10 seul ($\geq$1 code)	0.53	0.90	0.52	0.90	0.66	0.71	0.67 (0.58-0.77)	N/A
Baseline algo.	0.55	0.89	0.58	0.88	0.69	0.73	0.67 (0.58-0.76)	N/A
Baseline algo., plus contexte	0.64	0.76	0.58	0.81	0.72	0.69	0.68 (0.59-0.78)	N/A
Carroll's algo.	0.56	**0.98**	0.55	**0.98**	0.77	0.71	0.71 (0.64-0.80)	0.91 (0.86-0.95)
PheVis (setting a)	0.62	0.90	0.68	0.87	0.76	0.75	0.72 (0.63-0.82)	0.88 (0.82-0.93)
PheVis modifié (setting b)	0.68	0.88	0.77	0.83	0.78	**0.79**	**0.75** (0.66-0.85)	0.85 (0.78-0.92)
Carroll's algo. (Selon Carroll *et al.*[7])	**0.90**	N/A	0.65	N/A	N/A	N/A	N/A	**0.95**
PheVis (Selon Ferté *et al.*[5])	0.65	0.96	**0.94**	0.74	N/A	N/A	N/A	0.94

TABLE 1: Performances pour le phénotyping au niveau patient. PheVis setting a is ω = 10, half$-$life = 365 ; PheVis modifié setting b is ω = 2, half$-$life = 60. * intervals de confiance calculés par bootstrap.

Discussion Malgré des performances légèrement supérieures à des procédures simples de recherche de patient, les deux algorithmes testés presentes des performances similaire. Le gain de performance est à mettre en perspective du coût d'adaptation de ces algorithmes à un nouveau contexte. Le premier algorithme adapté depuis des hôpitaux américains présente une charge d'adaptation plus lourde, car il nécessite un travail manuel pour l'adaptation des règles d'extraction d'information. L'adaptation est indispensable pour obtenir des résultats satisfaisants. Cependant, il est moins gourmand en ressources informatiques que le second algorithme, semi-supervisé. Cet algorithme est cependant plus facilement extrapolable à d'autres environnements (autre clinique, autre langue) et d'autres pathologies.

Références

1. Katherine M Newton, Peggy L Peissig, Abel Ngo Kho, Suzette J Bielinski, Richard L Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J Kullo, Rongling Li, et al. Validation of electronic medical record-based phenotyping algorithms : results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*, 20(e1) :e147–e154, 2013.
2. Chunhua Weng, Nigam H Shah, and George Hripcsak. Deep phenotyping : embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of biomedical informatics*, 105 :103433, 2020.
3. William Digan, Aurélie Névéol, Antoine Neuraz, Maxime Wack, David Baudoin, Anita Burgun, and Bastien Rance. Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, 28(3) :504–515, 12 2020.
4. Robert J. Carroll, Anne E. Eyler, and Joshua C. Denny. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis, mar 2015.
5. Thomas Ferté, Sébastien Cossin, Thierry Schaeverbeke, Thomas Barnetche, Vianney Jouhet, and Boris P Hejblum. Automatic phenotyping of electronical health record : Phevis algorithm. *Journal of Biomedical Informatics*, 117 :103746, 2021.
6. Sebastien Cossin, Vianney Jouhet, Fleur Mougin, Gayo Diallo, and Frantz Thiessard. IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, 2125 :94, 2018.
7. Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, Elizabeth W Karlson, Raul G Perez, Vivian S Gainer, Shawn N Murphy, Eric M Ruderman, Richard M Pope, Robert M Plenge, Abel Ngo Kho, Katherine P Liao, and Joshua C Denny. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1) :e162–e169, 02 2012.