Evaluating the Portability of Rheumatoid Arthritis Phenotyping Algorithms: case study on French EHRs

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

Evaluating the Portability of Rheumatoid Arthritis Phenotyping Algorithms: case study on French EHRs

Thibaut FABACHER[a,b,c,d,1], Erik-André SAULEAU[a,b], Noémie LECLERC DU SABLON [a], Hugo BERGIER [a], Jacques-Eric GOTTENBERG[a], Adrien COULET[c,d]*, and Aurélie NÉVÉOL[e]*

[a] *University hospital of Strasbourg;* [b] *ICube Laboratory, Strasbourg, France;* [c] *Inria Paris, France;* [d] *Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, France;* [e] *Université Paris-Saclay, CNRS, LISN, Orsay, France; *Equal contribution*

Abstract. Previous work has successfully used machine learning and natural language processing for the phenotyping of Rheumatoid Arthritis (RA) patients in hospitals within the United States and France. Our goal is to evaluate the adaptability of RA phenotyping algorithms to a new hospital, both at the patient and encounter levels. Two algorithms are adapted and evaluated with a newly developed RA gold standard corpus, including annotations at the encounter level. The adapted algorithms offer comparably good performance for patient-level phenotyping on the new corpus (F1 0.68 to 0.82), but lower performance for encounter-level (F1 0.54). Regarding adaptation feasibility and cost, the first algorithm incurred a heavier adaptation burden because it required manual feature engineering. However, it is less computationally intensive than the second, semi-supervised, algorithm.

Keywords. Natural Language Processing, Phenotyping, Rheumatoid Arthritis

1. Introduction

Electronic Health Records (EHRs) enable secondary use of hospital data, and in particular the design and conduct of clinical studies. A first step of such studies is the definition of a cohort of patients who share a specific condition. This task is referred to as electronic phenotyping and is often more complex than a simple query [1,2]. Searching a unique phenotypic trait in EHRs usually requires covering both structured fields and unstructured texts in a specific time frame. Furthermore, phenotyping algorithms may not transfer well from one clinical setting to another because of variations in data collection, clinical practice, coding of medical acts, policies, or language used in clinical notes. In general, cohort definitions rely on phenotyping at the patient level, but a finer granularity is necessary in some cases, especially when monitoring chronic diseases, since a same patient's encounters (i.e., visits) may or may not be related to the phenotype.

In this work, we study the portability of phenotyping algorithms for Rheumatoid Arthritis (RA), a long-term autoimmune pathology that primarily affects joints. We explore with RA because it is relatively frequent, it poses clinical questions (e.g., what is a patient's prognosis, or best treatment options) and because phenotyping algorithms for RA have been described in the literature [3,4]. In particular, we compare three RA

[1] Corresponding Author: Thibaut FABACHER, thibaut.fabacher@chru-strasbourg.fr

phenotyping algorithms on unseen EHR data to address the following questions: Which one is the most efficient in terms of performance and speed? Which one is easier to adapt to a new hospital, here the University Hospital of Strasbourg (UHS), France? Which one is prone to performance decrease when transferred?

2. Materials and Methods

Records of patients with encounters in 2015-2020 and high probability of RA are extracted from the UHS health information system. Specifically, we select patients with at least one ICD-10 code related to RA and one reference to RA in a clinical text over this time period. ICD-10 codes for RA are M060*, M068*, M069*, M058*, M059*, M053*, M050*; and detection of RA in French texts is performed with the regex pol(i|y)arth?rites? *?rh?umat, searching for "polyarthrite rhumatoïde" and its variations due to typos. Clinical notes (discharge summaries, progress reports, etc.), diagnostic codes (ICD-10), drug prescriptions and laboratory results were extracted for these patients. We excluded encounters only associated with ICD-10 codes or prescriptions. We excluded notes with content limited to ICD-10 codes or antecedent. This study is listed on UHS study register and follows the hospital clinical study protocol.

Data are split in three patient sets. 11% are randomly selected to form the exploration set, which is used to evaluate regex from Carroll's algorithm. The remaining 89% is split in a customized way so 85% constitutes our train set and 4% our test set. The customized sampling strategy is performed to obtain balanced groups of patients in the test set. Both train and exploration sets are used to train the PheVis Algorithm. The test set is annotated and used to evaluate all different methods.

For the evaluation of phenotyping algorithms, we manually annotated our test set both at the patient and encounter levels. Accordingly, each encounter is annotated by two distinct individuals (among one rheumatologist and two public-health physicians). Each encounter is annotated with one of the following four labels: RA+ if the encounter is due to RA : diagnosis, assessment of disease progression, therapeutic management of the disease, management of complications of the disease; RA- if the encounter is not related to RA, even if the patient has an active RA; doubtful if the encounter cannot be confidently classified in relation to RA. To reach consensus, encounters annotated with two distinct labels are identified and discussed during a meeting. Doubtful encounters are ultimately labeled as RA- for method assessments, as the classification task that is evaluated is binary. For patient-level annotations, if a patient has at least one encounter labeled as RA+, she/he is labeled as RA+ at the patient-level, and as RA- otherwise.

Baseline Algorithm: Encounters are classified RA+ if they have at least one ICD-10 code for RA and at least one mention of RA in a clinical text. Matching with ICD-10 codes is performed according to the list of ICD-10 codes described in the Data Collection paragraph. Matching with clinical texts is performed with a dictionary-based NER tool, named IAMsystem [4]. In addition, we performed two filtering to avoid false positives, referred to as contextualization in the following. First filtering excludes parts of the text concerning medical history. To this aim, we use a house-made algorithm, for section segmentation. Second filtering consists in taking into account the context of RA mentions in clinical texts. To this aim, we use FastContext [5] and more specifically its implementation named IAMFastContext. With this tool, mentions of RA which are negated, hypothetical, historical or related to relatives or other persons are filtered out. **Carroll's Algorithm** [6,7] uses pretrained penalized logistic regression. Carroll et al.

provide parameters of the regression, to enable the reuse of the classifier on new data. This algorithm takes as input structured data (ICD-9 codes, drug prescriptions, lab results) and named entities found in clinical texts with a set of regex. To adapt this algorithm to the UHS, the ICD-9 codes are manually converted to ICD-10, drug prescriptions and laboratory results of the UHS are adapted to be consistent with the classifier, and finally, the set of regex provided in English is adapted to French. The original regex, their translations and adaptation to French are available at https://gitlab.inria.fr/heka/ra_phenotyping/. We apply Carroll's algorithm with coefficients provided in the original article, and a probability threshold of 0.5 for classification of RA patients. **PheVis Algorithm** [4] leverages the PheNorm [8] method to classify patients following a semi-supervised approach. It classifies not only at the patient level, but also at the encounter level. PheVis is a two-stage approach, as it relies first on the definition of a silver standard of automatically annotated examples, that is used, second, to train a supervised model. PheVis takes as input ICD-10 codes and UMLS entities automatically extracted from EHR narratives with NLP. Our adaptation of the PheVis algorithm relies on the IAMSystem for entity extraction and normalization [9] to ensure comparability with the original PheVis study [4]. To test the portability of PheVis to the UHS setting, we test the best hyperparameters reported by PheVis authors.

Evaluation metrics Phenotyping is assessed with precision, negative predictive value (NPV), specificity, recall (or sensitivity), balanced accuracy, accuracy, F1 score and Area Under the ROC Curve (AUC). Confidence intervals are computed using bootstrap. Experiments use R version 4.1 and a personal computer under Windows 10, with 64Gb of memory and an Intel(R) Xeon(R) CPU E3-1245 v5.

3. Results

We found 4,100 patients with at least one ICD-10 code for RA and one reference to RA in narratives. The 410 patients with the most recent first encounter were selected as a validation set for future work. Remaining 3,690 patients were split in 410 (11%), 3,140 (85%) and 140 (4%) patients to constitute exploration, train and test sets. These include 3,826, 33,007 and 1,552 distinct encounters with at least one clinical text.

Of the 1,552 encounters selected for manual annotation, after consensus on the annotation, 1,146 were classified as RA-, 358 as RA+ and 48 as doubtful. Inter-annotator agreement was substantial (Cohen's kappa = 0.80). At the patient level (n=140), 52 (37%) were classified as RA+ and 88 (63%) as RA-.

Table 1 and 2 summarize the results of our evaluation of phenotyping algorithms. Baseline Algorithm For hospital encounters classification, F1 was 0.60 [0.55-0.64] and 0.61 [0.55-0.65] respectively for the basic algorithm without and with contextualization. For patient classification, F1 was 0.69 [0.60-0.78] and 0.73 [0.65-0.83] without and with contextualization of named entities. Contextualizing NERs improves performance, in particular precision. Carroll's Algorithm For patient classification, F1 is 0.82 [0.75-0.90] . Results are similar to those of Carroll et al. [7] (AUC=0.94 vs. AUC=0.95), with a higher specificity (0.82 vs. 0.65) and a lower precision (0.75 vs. 0.90). Carroll's algorithm is not available for encounter-level phenotyping. Phevis Algorithm For patient classification, results are lower than those reported in Phevis paper (AUC=0.87 vs 0.94), F1 =0.68 [0.59-079]). F1 is lower, 0.54 [0.50-0.58] for encounter-level.

The baseline algorithm is fairly easy to implement. ICD-10 codes are easy to extract from structured data. Regex matching is also fast, taking less than 20 minutes in our

setting. Implementing the Carroll's algorithm took longer. About two working days was necessary to translate regex from English to French. It took one week to examine and modify regex with the exploration set. Searching to match all regex on the test set took about one hour. Implementing the logistic regression took half a day and the execution time of the logistic regression is almost instantaneous. The implementation of PheVis algorithm took more time. For data preparation, the NER with IAMsystem algorithm, took about two days to run on the exploration, train, and test datasets. Training a model took about 10 minutes. Once the classification algorithm is trained, application on new data is fast and takes about one minute.

Table 1. Performances for RA phenotyping at the patient level. PheVis setting is ω= 5, half-life = Inf

Methods	Prec.	NPV	Spe.	Rec.	bal Acc.	Acc.	F1	AUC
ICD-10 alone (≥1 code)	0.55	0.91	0.56	**0.90**	0.73	0.69	0.68 (0.58-0.77)	N/A
Baseline algo.	0.58	0.88	0.64	0.85	0.73	0.71	0.69 (0.60-0.78)	N/A
Baseline algo., plus context	0.67	0.87	0.76	0.81	0.77	0.78	0.73 (0.65-0.83)	N/A
Carroll's algo.	0.75	**0.94**	0.82	**0.90**	**0.84**	**0.85**	**0.82 (0.75-0.90)**	0.94 (0.89-0.99)
PheVis	0.62	0.84	0.72	0.77	0.73	0.74	0.68 (0.59-0.79)	0.87 (0.81-0.93)
Caroll, reported *et al.*	**0.90**	N/A	0.65	N/A	N/A	N/A	N/A	**0.95**
Phevis, reported *et al*	0.65	0.96	**0.94**	0.74	N/A	N/A	N/A	0.943

Table 2. Performances for RA phenotyping at the encounter level. PheVis setting is ω= 5, half-life = Inf

Methods	Prec.	NVP	Spe.	Rec.	bal Acc.	Acc.	F1	AUC
Baseline algo.	0.62	0.88	0.89	0.58	0.75	0.82	0.60 (0.55-0.64)	N/A
Baseline algo.	0.66	0.87	0.92	0.55	0.76	0.83	0.60 (0.55-0.64)	N/A
Baseline algo., plus context	**0.71**	0.87	**0.94**	0.53	**0.79**	**0.84**	**0.61 (0.56-0.65)**	N/A
PheVis	0.43	**0.89**	0.72	**0.71**	0.66	0.72	0.54 (0.50-0.58)	0.79 (0.76-0.82)

4. Discussion

Porting phenotyping algorithms from one setting to another remains a challenge. On UHS data, PheVis appears to have lower performance to Carroll's and baseline algorithms for patient phenotyping. Our adaptations of algorithms yield performance slightly lower than those reported in the literature (Table 1). The better results achieved so far may be due to the definition of RA+ patients, i.e. those with a history of RA or active RA. A recent study makes the same observation about the difficulty of adaptation and highlights the difficulty of defining the phenotyping task [10]. One originality of our study is the evaluation of algorithms at the encounter level. Although the authors of PheVis considered phenotyping encounters, their algorithm was evaluated only at the patient level. Our study suggests that PheVis is not superior to other algorithms at the encounter level. The rather good results we observed with the baseline algorithm, regarding what is reported in literature, may be attributed to an improvement of the coding in French hospitals.

For patient phenotyping, the majority of false positive predictions are due to our definition of RA+ patients. Majority of the false positives are patients with a history of RA. Contextualization improves precision by removing part of the patient's history, but redundancies between encounters are found even if they are not directly related to the chronic disease [11]. Reducing this should reduce the number of false positives.

In this initial study of RA phenotyping in French EHRs, our goal was to use state-of-the-art algorithms validated in previous studies on new patient data. For Carroll's algorithm, differences between languages in terms of sentence construction make it difficult to translate complex regex from one language to another. PheVis phenotyping performance is good at the patient level. Adjusting the hyperparameters to the local context and pathology would allow for even better results. Unsurprisingly, encounter-level performance is underwhelming, as it is a harder task. Results may be improved by using more complex unsupervised machine learning classifiers like Phe2vec [12].

The Phevis method requires as little expert knowledge as the rule-based algorithm, this method can be used across hospitals, provided a suitable NER algorithm is available. Carroll's algorithm is more difficult to adapt to another phenotype. Feature extraction could be easily adapted; however, the availability of an annotated dataset to train the classifier is a bottleneck.

5. Conclusion

The two algorithms tested for RA phenotyping are transferable to the context of the UHS. In both cases, adaptation required a significant amount of time, whether for the translation of regular expressions or the implementation of a NER algorithm. The performance gain compared to a baseline algorithm relying solely on ICD-10 codes is surprisingly low. Previous studies did not always consider the baseline in their evaluation and encounter-level phenotyping needs to be better considered. More advanced machine learning algorithms, taking into account redundancy or more specific silver standard, could improve performance in future work.

References

[1] Newton KM, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. JAMIA 2013. doi:10.1136/amiajnl-2012-000896.

[2] Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality-Towards scalability, portability, and interoperability. J Biomed Inform 2020. doi:10.1016/j.jbi.2020.103433.

[3] Carroll RJ, Eyler AE, Denny JC. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis. Expert Rev Clin Immunol 2015. doi:10.1586/1744666X.2015.1009895.

[4] Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform 2021. doi:10.1016/j.jbi.2021.103746.

[5] Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 2013;192:677–81.

[6] Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res 2010. doi:10.1002/acr.20184.

[7] Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. JAMIA 2012. doi:10.1136/amiajnl-2011-000583.

[8] Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, et al. Enabling phenotypic big data with PheNorm. J Am Med Inform Assoc JAMIA 2018;25:54–60. doi:10.1093/jamia/ocx111.

[9] Cossin S, Jouhet V, Mougin F, Diallo G, Thiessard F. IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates 2018. doi:10.48550/arXiv.1807.03674.

[10] Zheng HW, Ranganath VK, Perry LC, Chetrit DA, Criner KM, et al. Evaluation of an automated phenotyping algorithm for rheumatoid arthritis. J Biomed Inform 2022. doi:10.1016/j.jbi.2022.104214.

[11] Digan W, et al. Evaluating the Impact of Text Duplications on a Corpus of More than 600,000 Clinical Narratives in a French Hospital. MEDINFO 2019 2019:103–7. doi:10.3233/SHTI190192.

[12] De Freitas JK, Johnson KW, et al. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns 2021. doi:10.1016/j.patter.2021.100337.