Evaluating the Portability of Rheumatoid Arthritis Phenotyping Algorithms: case study on French EHRs at the Patient and Encounter Level

Anonymous ACL submission

Abstract

High-throughput phenotyping can accelerate the development of statistical analysis from cohorts of Electronic Health Records. 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 to the context of the new hospital and evaluated with a newly developed RA gold standard corpus, including annotations at the encounter level. The adapted algorithms offer comparable performance for patient-level phenotyping on the new corpus (F1 0.71 to 0.79), performance is lower for encounter-level phenotyping (F1 0.54 to 0.57), illustrating 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.

1 Introduction

Electronic Health Records (EHRs) enable secondary use of hospital historical data, and in particular the design and conduct of clinical studies. One of the first steps of such clinical studies is the definition of a cohort of patients who share a specific condition or outcome. This task is usually referred to as *electronic phenotyping* (or phenotyping) and is often more complex than a simple one-word query (Newton et al., 2013; Weng et al., 2020). One difficulty of cohort definition comes from the complex nature of EHR data, which include heterogeneous structured and unstructured data over long periods of time. This implies that searching a unique phenotypic trait may require a search both on structured fields and unstructured texts in a specific time frame. Another difficulty comes from the fact that phenotyping algorithms may not transfer well from one clinical setting to another. Indeed, variations in data collection, clinical practice, coding of medical acts, policies, languages cause that a locally-developed phenotyping algorithm may require significant adaptation to be transferred to a new clinical setting. In general, cohort definitions rely on phenotyping at the *patient level*, but a finer granularity may be necessary when physicians are interested in monitoring a specific, especially chronic, disease. They may need not only to know which patients have the phenotype, but also which *encounters* are related to the phenotype. Here, we define an encounter as a patient's visit to the hospital, whether as a hospital stay or an outpatient consult.

In this work, we particularly 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 a frequent pathology, it is associated with many current clinical questions that could be answered with clinical investigations (*e.g.*, predicting patient's prognosis, or best treatment options) and because some phenotyping algorithms for RA have been described in the literature (Carroll et al., 2015; Ferté et al., 2021). In particular we evaluate whether previous RA phenotyping algorithms can be easily and efficiently deployed in *Anonymous* hospital (AH). We also evaluate the performance of existing RA phenotyping algorithms to identify patients with RA, as well as specific encounters that are primarily associated with RA in the patient's clinical history.

Phenotyping algorithms rely generally on two steps (Alzoubi et al., 2019): First a data mart is created, through feature extraction from a clinical data warehouse and in turn classification algorithms are defined. While structured data can be easily analyzed for extraction, phenotyping algorithms also leverage Natural Language Processing (NLP) to identify important information that may only be

present in unstructured data. The NLP methods in-use include symbolic methods such as regular expressions (or regex) as well as statistical methods such as pre-trained deep learning models.

For classification algorithms, two main families of approaches can be distinguished (Shivade et al., 2014; Banda et al., 2018): Rule-based and statistical approaches. In the case of the rule-based approach, logical rules are applied to the EHRs to define the patient phenotype. The rules can be as simple as matching a simple pattern in narrative data. But they can also consider a complex combination of sources of information (ICD codes, concepts extracted from text, laboratory results, etc.)(Oake et al., 2017). The context of entities detected in text such as negation or temporality markers can also be used. Statistical approaches are broadly speaking based on machine learning methods, out of which two main categories can be distinguished: supervised and unsupervised methods. Supervised methods require pre-labeled data that is used to train a model. Unsupervised methods use unlabeled data, such as clustering methods.

Both types of approaches require, at some point, the incorporation of expert knowledge. In rule-based approaches, an expert classically defines the decision rules, whereas in statistical approaches the expert may manually label (or annotate) some data. Both tasks are time consuming. Lately, supervised methods appear to provide better results, but the time needed for establishing good quality labeled datasets can be huge. Rule-based approaches may seem easier to define, but are less accurate if too simple. Rule-based and statistical approaches can be combined to increase accuracy or limit annotation requirements. Semi-supervised learning is an example of such a hybrid approach. Rules are used to automatically pre-label a set of data, to constitute what is commonly named a *silver standard* (in comparison with manually pre-labeled data named *gold standard*) and then a supervised machine learning algorithm is trained on this silver standard.

Good phenotyping methods should offer good performance (they do not miss patients with the condition, and do not falsely identify patients without it) and they should be easy to adapt from one clinical setting to another (including settings where language or policies differ).

In this work, we consider and compare three different approaches that we tested on unseen EHR data. The first is a rule-based approach, hereafter referred as *baseline algorithm*, which relies only on ICD-10 diagnostic codes already present in EHRs and a Named Entity Recognition (NER) algorithm to capture the condition name in clinical texts. It is a standard baseline approach for phenotyping that requires little new expert knowledge and can be easily applied to a new clinical setting. In the case of RA, this algorithm previously showed low specificity and accuracy (Liao et al., 2010). The second is a supervised algorithm, hereafter referred as *Carroll's Algorithm*. It was first described by Liao *et al.* (Liao et al., 2010) and later tested for portability on three other hospitals by Carroll *et al.* (Carroll et al., 2012). The third algorithm, hereafter referred as *PheVis Algorithm*, is based on a semi-supervised approach (Ferté et al., 2021).

These three different algorithms can be adapted to new data. But these induce different human and machine costs. We sought to address the following research questions: Which algorithm is the most efficient in terms of performance and time? Which one is easier to adapt to a new hospital? Which one is prone to performance decrease when moved? Should the hospital encounter not be the granularity level for phenotyping? In particular for chronic diseases?

The portability study presented herein provides insight on these issues, in the context of deploying RA phenotyping algorithms within a French hospital.

The main contributions of this study are:

- a comparison of performance and necessary efforts when adapting the state-of-the-art algorithms for RA phenotyping, in the context of a French hospital;
- an evaluation of these algorithms for RA phenotyping at the encounter level; and
- an adaptation of a state-of-the-art algorithm for the detection of RA at the encounter level.

The next section presents the data, the different phenotype algorithms, their adaptation to our local setting and the method of evaluation. The Results Section reports the outcomes of our comparative portability study of phenotyping algorithms, at both patient and encounter levels. The article ends on a discussion about our results, the limits of the work, and a conclusion.

2 Materials and Methods

2.1 Data Collection

We use data from the *Anonymous* Hospital (AH) in France. Records for patients with encounters between 2015 and 2020 who may have 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 in this time period. ICD-10 codes for RA are `M060*, M068*, M069*, M058*, M059*, M053*, M050*` and detection in free text is performed with the regex `pol(i|y)arth?rites? *?rh?umat` for the detection of RA in French (namely "polyarthrite rhumatoïde") and its variations due to typos. All available data for these patients are extracted, what consists of: all clinical notes (discharge summaries, progress reports, etc.), diagnoses encoded with ICD-10 codes, drug prescriptions and laboratory results. We excluded encounters associated with only ICD-10 codes or drug prescriptions. We excluded narrative data where the label of the question, in EHR, referred only to ICD-10 codes or history.

The data use and research projects are listed on the hospital study register, according to the AH policy for internal research projects conducted by hospital staff for internal use. No nominative data is used, except for those that may appear in clinical texts. No pseudonymization tool is used as all the work is performed inside the hospital network.

2.2 Exploration, train and test sets

Data are split in three parts. 11% of patients are randomly selected to form the *exploration set*, which is used to evaluate a set of regex written for Carroll's algorithm. The remaining 89% of patients is not split randomly, but in a customized way so 85% constitutes our *train set* and 4% our *test set*. The customized sampling strategy is performed to select patients for our *test set*. Those are in part ($\sim$30%) randomly sampled from the patients with discordant classifications between baseline and Carroll's algorithm and in part ($\sim$70%) randomly selected with weights that force to respect proportions of Carroll's algorithm classifications (*i.e.*, 34% positive and 66% negative). This customized sampling is used to obtain more balanced groups of patients in the test set. Train and exploration sets are used to train PheVis Algorithm. Test set is annotated and used to evaluate all different methods.

2.3 Gold standard

For the evaluation of phenotyping algorithms, we manually annotated our test set, both at the patient and encounter levels. For encounter-level, each encounter is annotated twice. This double annotation is performed by three distinct individuals: one rheumatologist, specialist of RA, and two public-health physicians. Each encounter is annotated with one of the following four labels:

no text if no clinical text documents the hospital encounter. Indeed, selected patients have at least one clinical text that matches for RA, but some of their encounters may not be associated with any text; *RA+* if the encounter is due to RA, particularly by falling in one of these cases: 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. For instance, if the patient is admitted for appendicitis; *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. If no agreement can be reached between annotators, the encounter is labeled as *doubtful*. Encounters labeled as *no text* or *doubtful* are ultimately labeled as *RA-* for method assessments, as the classification task that is evaluated is defined as 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.

2.4 Baseline Algorithm

The baseline algorithm classifies RA patients as positive or negative using ICD-10 codes and the mention of RA in clinical texts.

More precisely, patients 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 during the same encounter. 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 (Cossin et al., 2018).

In addition to dictionary-based matching, we performed two filtering based on the context of entities, to avoid false positives. First filtering excludes clin-

ical text in concern with medical history. To this aim, we defined a simple regex to detect and exclude health questionnaires which labels contain a notion of medical history and we use a house-made algorithm, for section segmentation in complex medical reports. This algorithm uses the semi standardized section heading templates and excludes sections related to medical records. Second filtering consists in taking into account the context of RA mentions in clinical texts. To this aim, we use FastContext (Chapman et al., 2013) and more specifically its implementation named IAMFastContext available at https://github.com/scossin/IAMsystemFastContext. With this tool, mentions of RA which are negated, hypothetical, historical or related to relatives or other persons are filtered out. Both these filtering are referred to as *contextualization* in the following.

2.5 Carroll's Algorithm

This algorithm was first described by Liao *et al.* (Liao et al., 2010) and later described in detail to enable reproducibility and portability in Carroll *et al.* (Carroll et al., 2012).

This method is twofold. The first step is the definition by domain experts of a list of features associated with RA. These consist of both findable elements from structured data of EHRs (ICD-9 codes, drug prescriptions, lab data) and named entities findable in clinical texts. To find entities in texts, authors propose to use existing NER tools, but also share a set of regex, which are easier to transfer from one hospital to another than NER tools.

The second step is the use of pretrained penalized logistic regression. The authors provide parameters of the regression, to enable the reuse of the classification model on new data. To adapt this algorithm to AH data, the ICD-9 codes are manually converted to ICD-10, drug prescription and laboratory data are adapted to be consistent with local AH data, and finally, the set of regex in English is adapted to French.

This adaptation was conducted in two stages: the first stage is a basic translation from English to French, and the second stage is a modification of basic translations, driven by an exploration of how expressions match on a subset of the data. For the first stage, we translate English terms of regex into French using our own knowledge, the DeepL translator (https://www.deepl.com/) and Wikipedia. For specific medical terms, multilingual ontologies of the SIFR BioPortal are used (Jonquet, 2019).

For the second stage, we manually explored how regex were matching with a random sample of elements of text from our exploration set. To this aim we randomly sampled paragraphs (pieces of texts separated by two new lines), with a weighting strategy for paragraphs that matched one regex, or one medical term present in a regex. Higher weight was given to paragraphs with the most matches in order to increase the chance to find a paragraph that matches partially or completely one of the regex. Sampled paragraphs are in turn manually reviewed and missing or incorrect matches lead to manual modifications of the triggered regex.

The complete list of regex, translations and modifications is available on gitlab (URL not included for anonymization).

Once regex are translated and modified, we apply Carroll's algorithm with coefficients provided in the original article. We use a probability threshold of 0.5 to classify RA patients.

2.6 PheVis Algorithm

The PheVis algorithm (Ferté et al., 2021) leverages the method proposed in PheNorm (Yu et al., 2018) to classify patients according to phenotypes a following semi-supervised approach. Accordingly, it presents the advantage of not requiring a large set of expert-labeled examples for training. PheVis builds on PheNorm and enables classification 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 annotated examples, that is used in a second stage to train a supervised model.

PheVis uses ICD-10 codes and entities automatically extracted from EHR narratives and mapped to Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) using NLP. Accordingly, each patient encounter is associated with a set of ICD-10 codes and CUIs.

Our adaptation of the PheVis algorithm relies on the IAM system for entity extraction and normalization; we used the same extraction method as the original PheVis algorithm to increase comparability with the original PheVis study (Ferté et al., 2021).

In order to test the portability of PheVis to our hospital setting, we tested first the best hyperparam-

eters reported by PheVis authors, but also tested different hyperparameters for optimization, following a grid search strategy. In particular, parameters named *omega* and *half−life* are optimized.

Accordingly, PheVis is trained on our exploration and train sets and next evaluated on the test set. PheVis standard way of defining silver standard is based on a *cumulative surrogate*. This surrogate is a standardized sum of the number of ICD-10 and CUI codes found for each encounter. Over time, the surrogate is cumulated with previous encounters, with an exponentially decreasing accumulation depending on the *half−life* hyperparameter.

Encounters with higher surrogate values are defined as positive in the silver standard, and those with lower values are negative. If the half−life is defined as infinite, the surrogate is a simple sum of the number of codes found over the encounters. We evaluate whether an alternative silver standard can be defined.

2.7 Revised PheVis

If the half−life is set to infinite, as Phevis article advises for chronic diseases, the last encounter is more likely to be used in the silver standard, even if this one is not directly due to RA. For this reason, we propose an improvement of the definition of the surrogate described in the following section.

The quantitative value of the encounter stays defined the same way. But the encounter assignment to either positive or negative in the silver standard follows a different strategy. We classify all encounters with the cumulative surrogate, but at the patient level, we reassign (to either positive or negative) encounters that are more likely or unlikely to be due to RA, based on each encounter surrogate, instead of the cumulative surrogate. For example, if the last two encounters of a patient are considered the two most likely to be positive among all encounters, the silver standard is defined as positive for the two encounters for this patient with the greater surrogate, regardless of cumulative surrogate.

2.8 Evaluation metrics

To evaluate the performance of the different phenotyping algorithms, the following metrics are used: precision (or positive predictive value, PPV), negative predictive value (NPV), specificity, recall (or sensitivity), balanced accuracy, accuracy, F1 score and Area Under the ROC Curve (AUC). Due to the unbalanced distribution of labels, the F1 score is used to determine the best algorithm. Confidence intervals are computed using bootstrap.

2.9 Technical set-up

Experiments are done with R version 4.1, with the PheVis package, Java IAMsystem and IAMsystem-FastContext, and performed on a personal computer under Windows 10, with 64Gb of memory and an Intel(R) Xeon(R) CPU E3-1245 v5.

3 Results

3.1 Data collection

We found 4,100 patients with at least one ICD-10 code for RA and one reference to RA in narratives of their EHRs, between 2015 and 2020 at AH. We excluded 410 patients with the most recent first encounter at the hospital to provide a validation dataset for future work.

Remaining 3,690 patients were split in 410 (11%), 3140 (85%) and 140 (4%) patients to constitute our exploration, train and test sets, respectively. These include 3,826, 33,007 and 1,668 distinct encounters with at least one clinical text, respectively.

3.2 Gold standard

Of the 1,668 encounters selected for manual annotation, 89 were classified as *no text*, and 1,579 were found in the healthcare software with sufficient narrative data. Of these, after consensus on the annotation, 1,172 were classified as *RA-*, 359 as *RA+* and 48 were classified as *doubtful* with the available data. Inter-annotator agreement (Viera and Garrett, 2005) was substantial, as Cohen's kappa coefficient was 0.80. When considering *doubtful* encounters as *RA-*, Cohen's kappa is 0.83. At the patient level, among the 140 annotated patients, 52 (37%) were classified as *RA+* and 88 (63%) as *RA-*.

Table 1 and 2 summarize the results of our comparative evaluation of phenotyping algorithms. Following sections detail some of these results.

3.3 Baseline Algorithm

For the classification of hospital encounters, F1 score was 0.59 [0.55-0.64] for the baseline algorithm without contextualization of the named entities, and 0.60 [0.56-0.64] with contextualization. For patient classification, F1 score was 0.67 [0.58-0.76] and 0.68 [0.59-0.78] without and with contextualization of the extracted name entities, respectively. The contextualization of extracted entities

Methods	Prec.	NPV	Spe.	Rec.	bal Acc.	Acc.	F1*	AUC*
ICD-10 alone ($\geq$1 code)	0.53	0.90	0.90	0.52	0.66	0.71	0.67 (0.58-0.77)	N/A
Baseline algo.	0.55	0.89	0.88	0.58	0.69	0.73	0.67 (0.58-0.76)	N/A
Baseline algo., plus context	0.64	0.76	0.58	0.81	0.72	0.69	0.68 (0.59-0.78)	N/A
Carroll's algo., non-modified regex	0.53	0.96	0.49	0.96	0.74	0.66	0.68 (0.60-0.77)	0.90 (0.84-0.95)
Carroll's algo., modified regex	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, revised (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. (as reported in Carroll *et al.*(Carroll et al., 2012))	0.90	N/A	0.65	N/A	N/A	N/A	N/A	**0.95**
PheVis (as reported in Ferté *et al.*(Ferté et al., 2021))	0.65	0.96	**0.94**	0.74	N/A	N/A	N/A	0.943

Table 1: Performances for RA phenotyping at the patient level. PheVis setting a is ω = 10, half$-$life = 365; Revised PheVis setting b is ω = 2, half$-$life = 60. * confidence interval calculated with bootstrap.ICD-10 alone is a reference to show patients who have at least one compatible ICD-10.

Methods	Prec.	NVP	Spe.	Rec.	bal Acc.	Acc.	F1*	AUC*
ICD-10 alone ($\geq$1 code)	0.59	0.88	0.60	0.88	0.82	0.74	0.59 (0.55-0.64)	N/A
Baseline algo.	0.63	0.87	0.55	0.90	0.83	0.73	0.59 (0.55-0.64)	N/A
Baseline algo., plus context	0.57	0.9	0.63	0.88	0.83	0.75	0.60 (0.56-0.64)	N/A
PheVis (setting a)	0.43	0.90	0.72	0.72	0.67	0.72	**0.54** (0.53-0.61)	0.82 (0.79-0.84)
PheVis, revised (setting b)	0.48	0.90	0.78	0.70	0.69	0.76	**0.57** (0.54-0.62)	0.82 (0.79-0.84)

Table 2: Performances for RA phenotyping at the encounter level. PheVis setting a is ω = 10, half$-$life = 365; Revised PheVis setting b is ω = 2, half$-$life = 60. * confidence interval calculated with bootstrap. Carroll's algorithm is not applied at encounter level. ICD-10 alone is a reference to show which encounters have at least one compatible ICD-10.

led to a better precision (0.64 *vs.* 0.55) but to a lower sensitivity (0.58 *vs.* 0.88).

3.4 Carroll's Algorithm

From 67 English regex in Carroll *et al.*, we generated 66 regex in French, which were in turn tested on 1,509 paragraphs, randomly sampled from the exploration set, for modification. Before modification, regex performances were: F1=0.84, precision=0.85 and recall=0.84. After modifications, performances went to: F1=0.89, precision=0.87 and recall=0.90.

For patient classification, Carroll's algorithm results are better with modified regex (F1 0.68 [0.60-0.77] *vs* 0.71 [0.64-0.80]). Results are lower than those reported in Carroll's paper (Carroll et al., 2012) (AUC=0.91 *vs.* AUC=0.95), a lower specificity (0.55 *vs.* 0.65) and a lower precision (0.56 *vs.* 0.90).

Carroll's algorithm is not available for encounter-level phenotyping and accordingly was not adapted to finer granularity.

3.5 Phevis Algorithm

For patient classification, hyperparameter tuning with grid search technique on all narratives and ICD-10 data reveals different best parameters to those reported in PheVis article: $\omega = 10$ and $half\text{-}life = 365$ result in the best predictions (F1 Score =0.72 [0.63-0.82]) for patient-level classification. For encounter-level classification, the best parameters are $half\text{-}life = 365$ and $\omega = 10$. F1 score was smaller for encounter-level classification 0.54 [0.53-0.61]. With the revised PheVis (to define an alternative silver standard) we gain a little in performance for patient-level classification (F1 score = 0.75 [0.66-0.85]), and for encounter-level classification (F1 score = 0.57 [0.57-0.65]).

3.6 The cost of adaptation

The baseline algorithm is fairly easy to implement. ICD-10 codes are easy to extract from structured data. Searching for regex matching is also fast, taking less than 20 minutes in our setting. Implementing the Carroll's algorithm took longer. About two working days (8 hours each) were 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 take about one minute. The annotation of our test set by one person (1,668 encounters) took about 27 hours.

4 Discussion

4.1 Phenotyping performance

Porting phenotyping algorithms from one setting to another remains a challenge. On AH data, PheVis appears to have slightly superior performance to Carroll's and baseline algorithm for patient phenotyping. Our adaptations of Carroll's and PheVis algorithms yield performance slightly lower than those reported in the literature (see the last two lines of Table 1). For Carroll's algorithm, this may be due to the level of the probability threshold at the end of the regression that needs to be reached to classify an example as *RA+*. Carroll *et al.* chose a threshold that yields a specificity of 97%, whereas we set the threshold to 0.5. We tested this setting with our data, but to reach a specificity of 97%, all patients were classified as *RA-*. One other possible explanation for this difference is the definition of the starting cohort. We limited the starting cohort to patients with at least one ICD-10 code related to RA and a reference to RA in a clinical text, and they used all patients available in their EHRs.
For PheVis, an explanation of the lower performance could be the duration of the follow-up. We limited our analysis to a period of five years, whereas a period of ten years is used in the original study. This longer follow-up period may lead to a richer silver-standard and to higher performance, due to the larger number of data per patient.

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.

There are some possible explanations for this observation. In France, ICD-10 codes are used to code hospital stays (inpatient), but not to code medical appointments (outpatient). As a result, these appointments may be more difficult to classify in the French setting since some of the features are missing. The rather good results we observed with the baseline algorithm, in regards to what is reported in the literature may be attributed to an improvement of the coding in French hospitals. ICD-10 coding in French hospitals is mainly used to evaluate hospital activity and to adapt public hospital funding. The use of these billing codes in clinical research was initially criticized because of the biases associated with the granularity and methodology of coding (Boudemaghe and Belhadj, 2017). However, over time, hospitals have strived to improve coding. In particular, the coding activity at the AH is now submitted to a stringent quality control process and has consequently improved.

4.2 Error analysis

For patient phenotyping, the majority of false positive predictions made by PheVis algorithm are due to our definition of *RA+* patients. We defined them as having at least one RA-related encounter during the follow-up period. PheVis algorithm tends to classify patients with a history of RA as *RA+*, even if there is no encounter directly related to the disease. For encounter level phenotyping, the main weakness of the algorithms is the precision: many encounters are falsely classified as *RA+*. Analysis of the medical record shows that patient's medical data often contain some text duplicated from previous records, even if the encounter is not directly related to the chronic disease (Digan et al., 2019). Removing duplicated text coming from previous encounters should reduce the number of false positives.

4.3 Choice of Algorithm

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. This allowed us to characterize the adaptation burden of a rule-based and machine learning algorithms. The baseline shows relatively competitive results in comparison with more complex algorithms, in particular at the encounter level. In the literature, a superiority of new and more complex methods is often observed. But if studies compare the new method they introduce with the latest complex one, they do not necessarily compare to baseline methods. This is an interesting direction for future work, which could assess whether a deep learning algorithm such as transformers or convolutional neural network would yield higher performance. However, one of the major difficulties remains the definition of the silver standard, which is not solved by more recent methods.

For Carroll's algorithm, differences between languages in terms of sentence construction make it difficult to translate complex regex from one language to another. We observed that the manual exploration and evaluation of regex and their subsequent modifications increased the performance of Carroll's algorithm (F1=0.71 *vs.* 0.68). This result suggests there is an added value in the manual regex engineering process.

PheVis algorithm presents the best result for phenotyping at the patient level. Unsurprisingly, results at the encounter level are underwhelming. Phenotyping at the encounter level is a harder task than at patient level. In the PheVis study, authors illustrate this difficulty with their attempt i.e., to phenotype an acute disease (tuberculosis), which can be compared to phenotyping at the encounter level, since acute diseases may be associated with a single encounter. A key issue is that at the encounter level, a large unbalance is usually observed between positive and negative, biasing classification tasks. To reduce this unbalance, a potential solution would be to split encounter-level phenotyping in two: First, the patient-level phenotyping and second, for positive patients only, the encounter-level phenotyping.

4.4 Named Entity Recognition

Features extracted from narrative text are key for phenotyping. The methods considered in this study use NLP to extract information from clinical text. Regular expressions can be effective for extracting basic information but they lack the ability to interpret context (*e.g.,* negations, history, hypotheses, etc.) adequately. Carroll's algorithm includes complex regex to account for the structure of English sentences. For example, to detect the positive mention of anti-CCP antibodies, five different regex are defined.The French language is very different, and led in our case to even more complex regex. An alternative to take context into account is to combine NER with a context detection algorithm based on trigger terms before and after entity mentions, or with deep learning methods that consider sentence-level contexts. These methods are more adaptable from one country to another if resources (*e.g.,* a list of trigger terms, a training set) are available for the language. Furthermore, these methods are generic enough to be transferable from one disease to another. For the baseline algorithm, adding context to NER did not impact the F1 score, but did increase the precision (0.64 *vs.* 0.55) and recall (0.81 *vs.* 0.58).

One challenge we faced is the widespread use of abbreviations and ambiguity in clinical narratives. One example is the use of "LES" standing for *Lupus érythémateux systémique* (Systemic lupus erythematosus or SLE in French). But in French, "les" is also the definite article "the" in English. Extraction of these abbreviations remains impossible with a simple regex or dictionary-based NER. Disambiguation of these terms, in particular with language representation learned with LSTM or attention-based architectures is a valuable direction for this task.

4.5 Portability

The Phevis method requires as little expert knowledge as the rule-based algorithm, since the expert only had to define main concepts from UMLS and ICD-10 codes associated with the disease. With a NER algorithm available, this method can be used in all hospitals and is not language specific. Carroll's algorithm is more difficult to adapt to another phenotype. Although features can be extracted with other tools than regex, the classifier is built with annotated data. To transfer this method, a pre-annotated dataset must be created.

4.6 Granularity

The choice of a method depends on the task at hand. To build up a cohort of patients with a specific chronic disease, a naive rule-based approach should be sufficient. To provide follow-up care for patients with a specific condition between two dates, the classical PheVis method may be more accurate. For encounter phenotyping, more complex algorithms may achieve better result.

5 Conclusion

The two algorithms tested for RA phenotyping are transferable to the context of the *Anonymous* Hospital. 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 considered the baseline in their evaluation and encounter-level phenotyping needs to be better considered.

References

Hadeel Alzoubi, Raid Alzubi, Naeem Ramzan, Daune West, Tawfik Al-Hadhrami, and Mamoun Alazab. 2019. A review of automatic phenotyping approaches using electronic health records. *Electronics*, 8(11).

Juan M Banda, Martin Seneviratne, Tina Hernandez-Boussard, and Nigam H Shah. 2018. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science*, 1:53–68.

Thierry Boudemaghe and Ihssen Belhadj. 2017. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). *International Journal of Epidemiology*, 46(2):392–392d.

Robert J. Carroll, Anne E. Eyler, and Joshua C. Denny. 2015. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis.

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. 2012. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169.

Wendy W. Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Mike Conway, Melissa Tharp, Danielle L. Mowery, and Louise Deleger. 2013. Extending the negex lexicon for multiple languages. *Studies in health technology and informatics*, 192:677–681. 23920642[pmid].

Sebastien Cossin, Vianney Jouhet, Fleur Mougin, Gayo Diallo, and Frantz Thiessard. 2018. 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.

William Digan, Maxime Wack, Vincent Looten, Antoine Neuraz, Anita Burgun, and Bastien Rance. 2019. Evaluating the impact of text duplications on a corpus of more than 600,000 clinical narratives in a French Hospital. In *medinfo 2019*, Lyon, France.

Thomas Ferté, Sébastien Cossin, Thierry Schaeverbeke, Thomas Barnetche, Vianney Jouhet, and Boris P Hejblum. 2021. Automatic phenotyping of electronical health record: Phevis algorithm. *Journal of Biomedical Informatics*, 117:103746.

Clement Jonquet. 2019. Semantic Indexing of French Biomedical Data Resources. In *Project Repository Journal*, volume 3, pages 16–19.

Katherine P Liao, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-treitler, Soumya Raychaudhuri, Peter Szolovits, Susanne Churchill, Shawn Murphy, Isaac Kohane, et al. 2010. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*, 62(8):1120–1127.

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. 2013. 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.

Justin Oake, Erfan Aref-Eshghi, Marshall Godwin, Kayla Collins, Kris Aubrey-Bassler, Pauline Duke, Masoud Mahdavian, and Shabnam Asghari. 2017. Using electronic medical record to identify patients with dyslipidemia in primary care settings: International classification of disease code matters from one region to a national database. *Biomedical Informatics Insights*, 9:1178222616685880. PMID: 28469428.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

Anthony J Viera and Joanne M Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

Chunhua Weng, Nigam H Shah, and George Hripcsak. 2020. Deep phenotyping: embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of biomedical informatics*, 105:103433.

Sheng Yu, Yumeng Ma, Jessica Gronsbell, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, et al. 2018. Enabling phenotypic big data with phenorm. *Journal of the American Medical Informatics Association*, 25(1):54–60.