Automatic phenotyping of electronical health record: PheVis algorithm

Title: Automatic phenotyping of electronical health record: PheVis algorithm

Corresponding author: Thomas Ferté, Université de Bordeaux, 146 Rue Léo Saignat 33076 Bordeaux, thomas.ferte@u-bordeaux.fr

Authors: Thomas Ferté[1,2], Sébastien Cossin[1,3], Thierry Schaeverbeke[4], Thomas Barnetche[4], Vianney Jouhet[1,3] and Boris P Hejblum[2].

1: Bordeaux Hospital University Center, Pôle de santé publique, Service d'information médicale, Unité Informatique et Archivistique Médicales, F-33000 Bordeaux, France

2: Univ. Bordeaux ISPED, Inserm Bordeaux Population Health Research Center UMR 1219, Inria BSO, team SISTM, F-33000 Bordeaux, France

3: Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France

4: Rheumatology department, FHU ACRONIM, Bordeaux University Hospital, F-33076 Bordeaux, France

Keywords (5): electronic health records; high-throughput phenotyping; phenotypic big data; precision medicine

Word count: 1969

ABSTRACT

Electronic Health Records (EHRs) often lack reliable annotation of patient medical conditions. Yu *et al.* recently proposed *PheNorm*, an automated unsupervised algorithm to identify patient medical conditions from EHR data. *PheVis* extends *PheNorm* at the visit resolution. *PheVis* combines diagnosis codes together with medical concepts extracted from medical notes, incorporating past history in a machine learning approach to provide an interpretable "white box" predictor of the occurrence probability for a given medical condition at each visit. *PheVis* is applied to two real-world use-cases using the datawarehouse of the University Hospital of Bordeaux: i) rheumatoid arthritis, a chronic condition; ii) tuberculosis, an acute condition (cross-validated AUROC were respectively 0.948 [0.945 ; 0.950] and 0.987 [0.983 ; 0.990]). *PheVis* performs well for chronic conditions, though absence of exclusion of past medical history by natural language processing tools limits its performance in French for acute conditions.

INTRODUCTION

As the amount of data collected on a daily basis from hospital health care system keeps increasing,[1] the appeal for leveraging the full potential of these data for research purposes and to investigate clinical questions is also becoming stronger than ever.[2–5] Yet, EHR data are quite different from research oriented data (e.g. cohort or trial data): i) they are less structured, more heterogeneous, ii) they present finer granularity, iii) data collection is done for health care purpose.[1,6–8] Currently, one of the main barriers to use such data for studying disease risk factors is the necessity to first identify patients having diseases of interest, a task that we will denote as phenotyping.

Several approaches have been recently proposed to phenotype patients.[9] They often rely on either rule-based algorithms specifically designed with clinicians, or on supervised models trained on annotated patient datasets. Such algorithms are limited because their development is disease specific, must be (re-)started from scratch for every new disease and demands a lot of clinician expertise time. In addition, portability and generalization to new databases (e.g. different hospitals) can often fail, requiring once again the process to be reiterated in the new institution. Hripcsak and Albers defined high-throughput phenotyping as an approach that "should generate thousands of phenotypes with minimal human intervention".[8] In the same vein, Yu *et al.* developed an unsupervised framework to phenotype diseases using EHR at the patient level.[10] It is fully automated and does not require any chart review neither rules definitions. They first apply a Surrogate-Assisted Feature Extraction (*SAFE*),[11] (an unsupervised feature selection method), and then *PheNorm*,[10] (an unsupervised phenotyping method), to estimate the disease status at the patient level. Both *PheNorm* and *SAFE* primarily rely on the main International Classification of Diseases (ICD) codes and the main Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) referring to the medical condition of interest (e.g. for rheumatoid arthritis: ICD codes *M05 Rheumatoid arthritis with rheumatoid factor* and *M06 Other rheumatoid arthritis,* and CUI *C0003873 Rheumatoid arthritis*). ICD codes come from the medical billing process of the hospital whereas UMLS codes are extracted from patient's clinical narrative notes with a natural language processing (NLP) tool. *SAFE* algorithm selects medical candidate features from a pool of publicly available

sources (e.g. Wikipedia, MedLine, etc). The main ICD codes and main CUI counts are used to create silver-standard labels for the medical condition of interest. Relevant features are then selected using an elastic-net regression on the silver labels. *PheNorm* algorithm also uses the main counts as a surrogate of the condition of interest and combines features into a phenotyping score through a noising-denoising step. The AUROCs of *PheNorm* were similar to supervised algorithms for rheumatoid arthritis.[10]

Although this framework is appealing, we need to go further than phenotyping at the patient level, especially for studying acute diseases (that can occur repeatedly) or for answering epidemiological questions (where temporal sequence is important). Phenotyping at the visit level allows taking into account the dynamic evolution of patient's conditions. In addition, the *SAFE-PheNorm* framework was developed using English databases, leveraging advanced NLP tools and rich terminologies available in English.[12,13] Portability to other languages is not straightforward, as they still often lack resources of matching quality.

We propose a new approach of unsupervised algorithm extending *PheNorm* at the visit level: *PheVis*. It will be evaluated for rheumatoid arthritis (RA) and tuberculosis (TB), a chronic and an acute condition respectively, using French EHRs from the University Hospital of Bordeaux.

METHODS

PheVis combines ICD billing codes together with medical concepts extracted from medical notes, incorporating past information through a user-tunable exponential decay. This creates a silver-standard surrogate of the medical condition of interest. Then variable selection (through elastic-net logistic regression) and pseudo-labellisation (using random-forest) are performed, leveraging extreme values of this silver-standard. Finally, a logistic regression model is estimated on those noisy labels to provide an interpretable "white box" predictor of the occurrence probability for a given medical condition at each visit. The different steps of *PheVis* are outlined in Figure 1. We will briefly explain the four main steps for training the *PheVis* algorithm (a detailed description of the method is available in the supplementary material).

Medical concepts of interest for a given disease are extracted from clinical notes and from ICD10 billing codes into a matrix where each row is a visit.[14] The first step is to sum main ICD codes (i.e. for RA: *M05 Rheumatoid arthritis with rheumatoid factor* and M06 *Other rheumatoid arthritis)* and CUI concepts (i.e. for RA: *C0003873 Rheumatoid arthritis*). As CUIs occurrences largely outnumber ICD ones, both are first standardized (otherwise the CUIs information would largely dominate the learning). Second, the information is cumulated over time according to an exponential decay law whose half-life can be tuned by the user through the λ parameter mentioned in figure 1 (for a chronic disease, all past history will be cumulated with an infinite half-life, while for an acute disease it will be set closer to 0 as the disease won't last for many visits). For instance a patient having RA might come for an acute infectious disease where clinicians won't focus on RA; past information is needed to predict current medical condition. Third, the previously defined cumulative sum $mCumul_{ij}$ for patient i at visit j is transformed into S_{ij}, a categorical variable that has three possible values: 0 (extremely low cumulative sum), 0.5 (medium cumulative sum), or 1 (extremely high cumulative sum), defining silver-standard labels. To set the thresholds separating these three categories, we use the prevalence of the main ICD code denoted p_{ICD}. S_{ij} is set to 1 if the $mCumul_{ij}$ is among the top $p_{ICD}/2$ visits, or set to 0 if $mCumul_{ij}$ is among the lowest $p_{ICD}/2$; other visits are set to 0.5. This takes into account the variability of disease prevalence in the training cohort. Fourth, the disease probability is estimated by i) first attributing a pseudo-value to each visit regarding the phenotype (either 0 or 1) using a random-forest trained on extreme visits with $S_{ij} = \{0,1\}$ (this optional step improves the performance by smoothing the predictor) ii) second fitting a logistic regression for estimating the probability of the medical condition of interest at each visit considered. This finally yields a phenotype prediction at the visit level.

RESULTS

Application design

We illustrate the *PheVis* method on RA, a chronic disease which cannot be cured, and active TB, an acute disease which usually last between 6 to 12 months.[15–18] The model performance was evaluated on an

imperfect gold standard for both diseases: for RA we used the presence of at least one rheumatoid arthritis form specifically used by rheumatologists at the University Hospital of Bordeaux in usual care, and for tuberculosis we manually reviewed patients with at least one mention of tuberculosis treatment while other patients were considered not having the disease. Latent tuberculosis was labelled as tuberculosis negative because even if it the bacterium is the same, symptoms, diagnosis and treatment are different. Patients were included if they had been hospitalized at the University Hospital of Bordeaux at least once since 2010 and if they had either one primary or secondary ICD code of RA (M05 or M06), or one biology measurement of Anti-Citrullinated Peptide Antibody. The cohort was split into training and test datasets at patient level with a 70% to 30% ratio. The cohort is described in Table 1, highlighting the discrepancy between ICD, CUI and gold-standard justifying the need for phenotyping algorithms.

TABLE 1 COHORT DESCRIPTION OF PHENOTYPING COHORT. *UNIVERSITY HOSPITAL OF BORDEAUX*.

	Test set		**Train set**	
	Visits	**Patients**	**Visits**	**Patients**
n	62105	2359	238156	9102
Gold standard PR (%)	7886 (12.7)	274 (11.6)	27087 (11.4)	953 (10.5)
ICD RA[1] ≥ 1 (%)	5823 (9.4)	901 (38.2)	21450 (9.0)	3683 (40.5)
CUI RA ≥ 1 (%)	8634 (13.9)	952 (40.4)	32782 (13.8)	3703 (40.7)
Gold standard TB (%)	90 (0.1)	5 (0.2)	618 (0.3)	49 (0.5)
ICD TB[2] ≥ 1 (%)	50 (0.1)	15 (0.6)	277 (0.1)	88 (1.0)
CUI TB ≥ 1 (%)	439 (0.7)	147 (6.2)	2394 (1.0)	647 (7.1)

1: ICD RA: M05, M06
2: ICD TB: A15, A16, A17, A18, A19

Four different prediction models were evaluated for each disease: i) using $mCumul_{ij}$ as a direct predictor of the phenotype, ii) our proposed *PheVis* approach, iii) a supervised elastic-net model trained using the gold standard directly, iii) a supervised random-forest trained using the gold standard directly. For *PheVis* we choose $\lambda_{RA} = \frac{ln(2)}{Inf} = 0$ and $\lambda_{TB} = \frac{ln(2)}{180}$ (tuberculosis typically lasting around 6 months).

Application results

Figure 2 shows individual *PheVis* predictions for four patients chosen for their various profiles. For each disease, the model was able to correctly capture the beginning of diseases. For TB however, it failed to return the prediction to a near zero probability when the disease ended. Two main reasons can explain this phenomenon: i) the cumulative constant λ value was poorly chosen; and ii) because of limitations in the NLP pre-processing step (no section segmentation to exclude past medical history), *PheVis* is not able to distinguish current information from past history. This has less impact for phenotyping RA because both current and past information positively predict the phenotype (which is generally the case for chronic conditions).

Figure 3 shows the global performance of *PheVis* on the test set compared to supervised models and $mCumul_{ij}$ for both diseases. The performances for RA were satisfying, while for TB only the Receiver Operating Characteristic (ROC) curve is satisfying (mostly an artifact due to the low prevalence of TB) and not the precision-recall one. For TB low performance is partly explained by the hard task of distinguishing active TB from latent TB and other mycobacteria infections as both have similar treatment and vocabulary.

Table 2 shows point performance for two arbitrary phenotyping decision rules: i) a predicted probability above 0.5 ii) a probability above the threshold maximizing the sum of the precision and the recall. Specificity and negative predictive values are good partly because the diseases are rare at the visit level. Matching the results from figure 2, the sensitivity/positive predictive values trade-off is better for RA than for TB.

TABLE 2 PHEVIS PERFORMANCE FOR DIFFERENT THRESHOLDS.

Disease	Threshold	SE	SP	PPV	NPV
Rheumatoid arthritis					
	0.5	0.81	0.91	0.58	0.97
	Optimal P-R* (0.87)	0.77	0.94	0.63	0.96
Tuberculosis					
	0.5	0.51	1.00	0.20	1.00

Optimal P-R* ($1.4\ 10^{-5}$)	0.99	0.95	0.03	1.00

* Threshold maximizing the precision recall sum.

DISCUSSION

We developed *PheVis* as an unsupervised automatic phenotyping algorithm at the visit level. It is able to achieve interesting performances for RA, which is promising for other chronic conditions, but suffers from limitations when it comes to acute conditions such as TB. *PheVis* is nearly fully automated, not requiring any (time-consuming and expensive) chart review, and its framework can in theory be used for different kind of medical conditions (either acute or chronic). It resembles the human medical probabilistic approach of diagnosis as the output is a probability taking into account the uncertainty of the information inside EHR.[19]

PheVis adds many innovations to the previous *PheNorm* algorithm it builds upon: the needs for standardizing the information from medical notes and ICD codes, the accumulation of past history with exponential decay, the definition of silver standard using ICD codes to take into account prevalence of the disease, and pseudo-labelling to improve performance and increase stability of predicted probabilities. Also we demonstrated the portability (and limitations) of those methods in French and in a different datawarehouse than the one used to develop *PheNorm*, with consistent performances for phenotyping RA.

These algorithms are highly sensitive to the input features, which emphasizes the needs for finer natural language processing tools able to perform semantic analysis. The use of other features such as biological test results or treatment should also be considered, as they should be highly predictive of the phenotype, but further works is needed to define how they could be integrated into the silver-standard surrogate.

The evaluation of the model is made against a questionable gold standard, mainly due to the lack of large annotated patient reference sets. For TB, the gold standard was manually curated, while for RA, we used a highly specific form but which might lack sensitivity: interestingly, upon manual inspection it appeared that *PheVis* was able to accurately recover RA patients visits of 5 patients who were not treated in the Rheumatology

department of the University Hospital of Bordeaux and thus had no record of this specific form, resulting in a failure of the gold standard. Such phenomena might underestimate the algorithm performance.

CONCLUSION

PheVis might be able to provide a probability for a large set of diseases and medical conditions with little effort. The performances might vary depending on the disease of interest, the database and the language. The use of those estimated probabilities opens new horizon for the use of EHR for medical and epidemiological research purposes.

SUPPLEMENTARY MATERIAL

ICD codes

TABLE S1MAIN ICD CODES OF RHEUMATOID ARTHRITIS AND TUBERCULOSIS USED BY PHEVIS.

Tuberculosis	A15, A16, A17, A18, A19
Rheumatoid Arthritis	M05, M06

Detailed methods

1. Input data

The input data of the *PheVis* workflow are the clinical notes and the ICD codes from a datawarehouse. All the notes and ICD codes are collapsed by visit, and a dictionary based named entity recognition is used to extract CUIs.[14] ICD codes are aggregated at one letter and two numbers level (i.e. M05.1 -> M05). This result in an X matrix of $\varphi \times P$ dimension where φ is the total number of visits and P the total number of ICD and UMLS concepts. We will denote $i \in \{1, \dots, n\}$ the patient index and $j \in \{1, \dots, v_i\}$ the visit index.

2. Build surrogate

As we do not have disease labels for the visits, we cannot use supervised modeling straight away. To be able to train our phenotyping algorithm, we first build a surrogate variable expected to be close to the true disease status. This surrogate is based on the main ICD and UMLS disease code.

We define mC_{ij} the standardized sum of main disease concepts as:

$$mC_{ij} = Z(mainICD_{ij}) + Z(mainCUI_{ij}) + min\left(Z(mainICD_{ij}) + Z(mainCUI_{ij})\right) \text{ and } Z(x) = \frac{x-\mu}{\sigma}$$

$mainICD$ and $mainCUI$ are main concepts related to the disease. For example for RA we used:

- $mainICD_{ij} = M05_{ij} + M06_{ij}$ with $M05_{ij}$ the number of times the code *M05 Rheumatoid arthritis with rheumatoid factor* was recorded for patient i at visit j, and similarly for $M06_{ij}$ and M06 *Other rheumatoid arthritis*
- $mainCUI_{ij} = C0003873_{ij}$ with *C0003873 Rheumatoid arthritis*

The standardization is necessary because CUIs occurrences largely outnumber ICD code numbers. Without it, the weight of ICD codes in the prediction would be negligible.

To phenotype a visit, it is necessary to take into account previous visits information. For example, a patient can be diagnosed RA at the age of 50, have a visit at 52 for an infectious event containing no information about RA. We want to be able to predict RA in both visits. To do so we propose to cumulate past history information with an exponential decay as follow:

$$mCumul_{ij} = mC_{ij} + mCumul_{ij-1}e^{-\lambda D_{ij}} \text{ with } mCumul_{i1} = mC_{i1} \text{ and } D_{ij} = t_{ij} - t_{ij-1}$$

λ is a constant parameter set by the user controlling the "loss of memory" of the algorithm. For easier interpretation one can prefer to set the value of half-life equals to $ln\,(2)/\lambda$. The half-life chosen was the usual duration of the disease (180 days for TB and infinity for RA).

The same exponential decay accumulation is applied to each ICD and UMLS codes. We also define five other variables:

- $lastvis_{ij} = mC_{ij-1}$
- $last5vis_{ij} = \sum_{h=j-5}^{j-1} mC_{ih}$
- $lastmonth_{ij} = \sum_{h=1}^{j-1} mC_{ih} \times \mathbf{1}_m$ with $\mathbf{1}_m = 1$ if $D_{ij} - D_{ip} \leq 30days, 0$ otherwise
- $lastyear_{ij} = \sum_{h=1}^{j-1} mC_{ih} \times \mathbf{1}_y$ with $\mathbf{1}_y = 1$ if $D_{ij} - D_{ip} \leq 365days, 0$ otherwise
- $Cum_{ij} = \sum_{1}^{j} mC_{ij}$

This yields an augmented matrix X^a of $\varphi \times (2P + 5)$ dimensions: CUIs and ICDs and their cumulated counts, and five new variables.

3. Variable selection

We used the *SAFE* algorithm to select predictive variables of interest and reduce the dimensionality of the optimization problem. First we used NLP to extract ICD and UMLS concepts in external resources.[15–18] A concept and its cumulative count were kept if it was found in the two resources. Then we categorized $mCumul_{ij}$ into $S_{ij} = \{0, 0.5, 1\}$. To define thresholds, we used $mainICD_{ij}$ which takes into account prevalence variability depending on the disease and the cohort. We define $quant_{extreme}$ as the quantile of visits having at least one main ICD code as:

$$quant_{extreme} = \frac{\#mainICD_{ij} \geq 1}{\varphi \times \omega} \text{ with } \omega \text{ a constant set to 2}$$

It allows us to define S_{ij} as:

$$S_{ij} = \begin{cases} 0, \text{ if visit belongs to } quant_{extreme} \text{visits with lower } mCumul_{ij} \\ 1, \text{ if visit belongs to } quant_{extreme} \text{visits with upper } mCumul_{ij} \\ 0.5 \text{ otherwise} \end{cases}$$

Then we trained a logistic regression with elastic-net penalization to select a subset X' of relevant variables from the X^amatrix:

$$S_{ij} \sim elasticnet\left(X_{ij}^a\right) \forall ij \text{ with } S_{ij} \in \{0,1\}$$

Of note, $mainICD$ and $mainCUI$ are always forced into the set of selected variables, while Cum_{ij} is systematically removed for acute conditions.

4. Pseudo-labelling

We attributed a pseudo-label $\{0,1\}$ to all visit. It increases the number of visits available to train the final logistic regression, and adds visits more uncertain phenotype status, which overall implies smoother predicted

probabilities and better performance. To perform this pseudo-labellisation, we train a random-forest with majority vote trees aggregation as:

$$S_{ij} \sim randomforest\left(X'_{ij}\right) \forall ij \text{ with } S_{ij} \in \{0,1\}$$

The model is used to predict $PL_{ij} = \{0,1\}$ status for each visit.

5. Probability estimation

To estimate the disease occurrence probability, we used a noising-denoising logistic regression with random intercept similarly to *PheNorm*. First, $max(10^5, \varphi)$ visits are randomly sampled with replacement with inverse probability weighting depending on PL_{ij} in order to balance the training set. This new matrix is denoted X^b. Then we performed a noising-denoising step to force the algorithm to use other variables than the main ICD and UMLS concepts (and thus avoid overfitting with respect to the surrogate). Every value of explanatory variables has a probability of $p_{bern} = 0.3$ to be replaced by the mean of the explanatory variable, and this noisy matrix is denoted X^n:

$$\begin{cases} \text{if } r_{bern_{ijc}} = 1 \text{ then } X^n{}_{ijp} = mean\left(X^b{}_p\right) \\ \text{if} r_{bern_{ijc}} = 0 \text{ then } X^n{}_{ijp} = X^b_{ijp} \end{cases} \quad \text{with } r_{bern} \sim Bern(p_{bern})$$

For the denoising step a logistic regression with random intercept is used:

$$logit\left(P\left(PL_{ij} = 1\right)\right) = X^{n^T}\beta + b_{0i} \text{ with } b_{0i} \sim N(0, \sigma_0^2)$$

And finally the probability of having the disease is estimated on the noise free matrix as:

$$P(Disease = 1) = expit(X^{bT}\beta)$$

COMPETING INTERESTS

None.

BIBLIOGRAPHY

1 Kim E, Rubinstein SM, Nead KT, *et al.* The Evolving Use of Electronic Health Records (EHR) for Research. *Semin Radiat Oncol* 2019;**29**:354–61. doi:10.1016/j.semradonc.2019.05.010

2 Beesley LJ, Salvatore M, Fritsche LG, *et al.* The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat Med* 2019;:1–28. doi:10.1002/sim.8445

3 Coorevits P, Sundgren M, Klein GO, *et al.* Electronic health records: new opportunities for clinical research. *J Intern Med* 2013;**274**:547–60. doi:10.1111/joim.12119

4 Danciu I, Cowan JD, Basford M, *et al.* Secondary use of clinical data: The Vanderbilt approach. *J Biomed Inform* 2014;**52**:28–35. doi:10.1016/j.jbi.2014.02.003

5 Sandhu E, Weinstein S, McKethan A, *et al.* Secondary Uses of Electronic Health Record Data: Benefits and Barriers. *Jt Comm J Qual Patient Saf* 2012;**38**:34–40. doi:10.1016/S1553-7250(12)38005-7

6 Wilcox AB. Leveraging Electronic Health Records for Phenotyping. In: Payne PRO, Embi PJ, eds. *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*. London: : Springer 2015. 61–74. doi:10.1007/978-1-4471-4646-9_4

7 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc JAMIA* 2013;**20**:144–51. doi:10.1136/amiajnl-2011-000681

8 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc JAMIA* 2013;**20**:117–21. doi:10.1136/amiajnl-2012-001145

9 Banda JM, Seneviratne M, Hernandez-Boussard T, *et al.* Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* 2018;**1**:53–68. doi:10.1146/annurev-biodatasci-080917-013315

10 Yu S, Ma Y, Gronsbell J, *et al.* Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc JAMIA* 2017;**25**:54–60. doi:10.1093/jamia/ocx111

11 Yu S, Chakrabortty A, Liao KP, *et al.* Surrogate-assisted feature extraction for high-throughput phenotyping. *J Am Med Inform Assoc JAMIA* 2017;**24**:e143–9. doi:10.1093/jamia/ocw135

12 Yu S, Cai T, Cai T. NILE: Fast Natural Language Processing for Electronic Health Records. *ArXiv13116063 Cs* Published Online First: 23 November 2013.http://arxiv.org/abs/1311.6063 (accessed 5 Sep 2019).

13 Bodenreider O, McCray AT. From French vocabulary to the Unified Medical Language System: A preliminary study. *Stud Health Technol Inform* 1998;**52**:670–4.

14 Cossin S, Jouhet V, Mougin F, *et al.* IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates. ;:7.

15 Polyarthrite rhumatoïde. http://www.lecofer.org/item-objectifs-0-19.php (accessed 12 Dec 2019).

16 Polyarthrite rhumatoïde. Wikipédia. 2019.https://fr.wikipedia.org/w/index.php?title=Polyarthrite_rhumato%C3%AFde&oldid=164700221 (accessed 12 Dec 2019).

17 Référentiel National de Pneumologie – CEP. http://cep.splf.fr/enseignement-du-deuxieme-cycle-dcem/referentiel-national-de-pneumologie/ (accessed 12 Dec 2019).

18 Tuberculose. Wikipédia. 2019.https://fr.wikipedia.org/w/index.php?title=Tuberculose&oldid=165260234 (accessed 12 Dec 2019).

19 Owens DK, Sox HC. Medical Decision-Making: Probabilistic Medical Reasoning. In: Shortliffe EH, Perreault LE, eds. *Medical Informatics*. New York, NY: : Springer New York 2001. 76–131. doi:10.1007/978-0-387-21721-5_3

FIGURES

Figure 1: Workflow of *PheVis*. Example: rheumatoid Arthritis (RA).

Figure 2: Individual prediction of rheumatoid arthritis and tuberculosis of *PheVis*. Each column corresponds to a disease, each row to a patient. Patient 1 has no disease. Patient 2 and 4 have rheumatoid arthritis and no tuberculosis. Patient 3 and 4 have tuberculosis.

Figure 3: Cross validated performance prediction of *PheVis* compared to supervised models and *mCumul* (Surrogate).

P	V	mainICD	mainCUI	$X_{1...P}$
1	1	0	5	...
1	2	0	40	...
1	3	1	2	...
2	1	0	0	...
...		...	...	...

1

Build surrogate

P	V	mC	mCumul	S	$X_{1...P}$
1	1	0.4	0.4	0.5	...
1	2	5.1	5.5	1	...
1	3	0.3	5.8	1	...
2	1	0	0	0	...
...		...	...	...	...

2

Pseudo-labellisation

P	V	S	PL	$X'_{1...P}$
1	1	0.5	1	...
1	2	1	1	...
1	3	1	1	...
2	1	0	0	...
...		...	...	...

3

Probability prediction

P	V	PL	Pr	$X'_{1...P}$
1	1	1	0.97	...
1	2	1	0.99	...
1	3	1	0.99	...
2	1	0	0.01	...
...		...	...	...

4

1

P: Patient ; **V**: Visit

mainICD: main ICD concept related to the disease.

mainCUI: main CUI concept related to the disease.

$X_{1...P}$ = {**ICDs**, **CUIs**}

2

mC: Standardised sum of main ICD and CUI (standardization is necessary otherwise CUI would dominate the learning).

$$\begin{aligned} &\mathrm{m}C_{ij} \\ &= \mathrm{Z}(mainICD_{ij}) + \mathrm{Z}(mainCUI_{ij}) \\ &+ min\left(\mathrm{Z}(mainICD_{ij}) + \mathrm{Z}(mainCUI_{ij})\right) \end{aligned}$$

$$Z(x) = \frac{x - \mu}{\sigma}$$

mCumul: Cumulate mC as rheumatoid arthritis is a chronic disease.

$$mCumul_{ij} = \mathrm{mC}_{ij} + mCumul_{ij-1} exp(-\lambda D_{ij})$$

λ set by the user depending on the disease duration (e.g $\lambda = 0$ for chronic diseases)

S: Silver standard label based on $mCumul$ and $mainICD$. The prevalence of the main ICD code is denoted p_{ICD}.

$$\begin{cases} \text{if } mCumul_{ij} \text{among top} \dfrac{p_{ICD}}{2} \text{ visits}, S_{ij} = 1 \\ \text{if } mCumul_{ij} \text{among lowest} \dfrac{p_{ICD}}{2} \text{ visits}, S_{ij} = 0 \\ \text{Otherwise } S_{ij} = 0.5 \end{cases}$$

3

$X'_{1...P}$: $\boldsymbol{X}'$ the matrix composed of all CUI and ICD features selected by SAFE (i.g. external ressources and elastic-net penalized logistic regression on S_{ij}) from the **X** matrix.

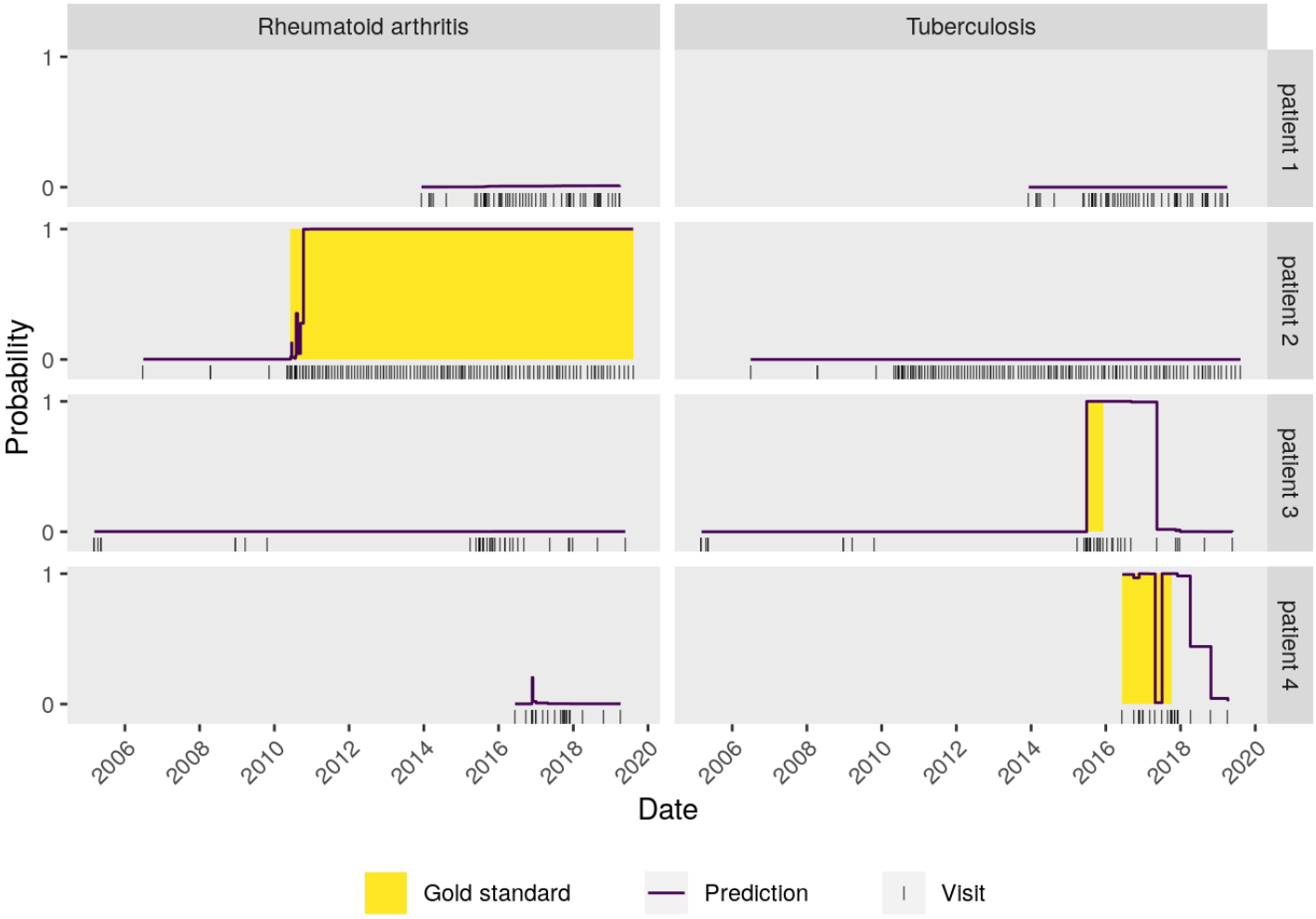
PL: Pseudo-labellisation assigns 0/1 to each visit, increase both performances and predictions stability. We denoted $\hat{h}$ the random forest predictor:

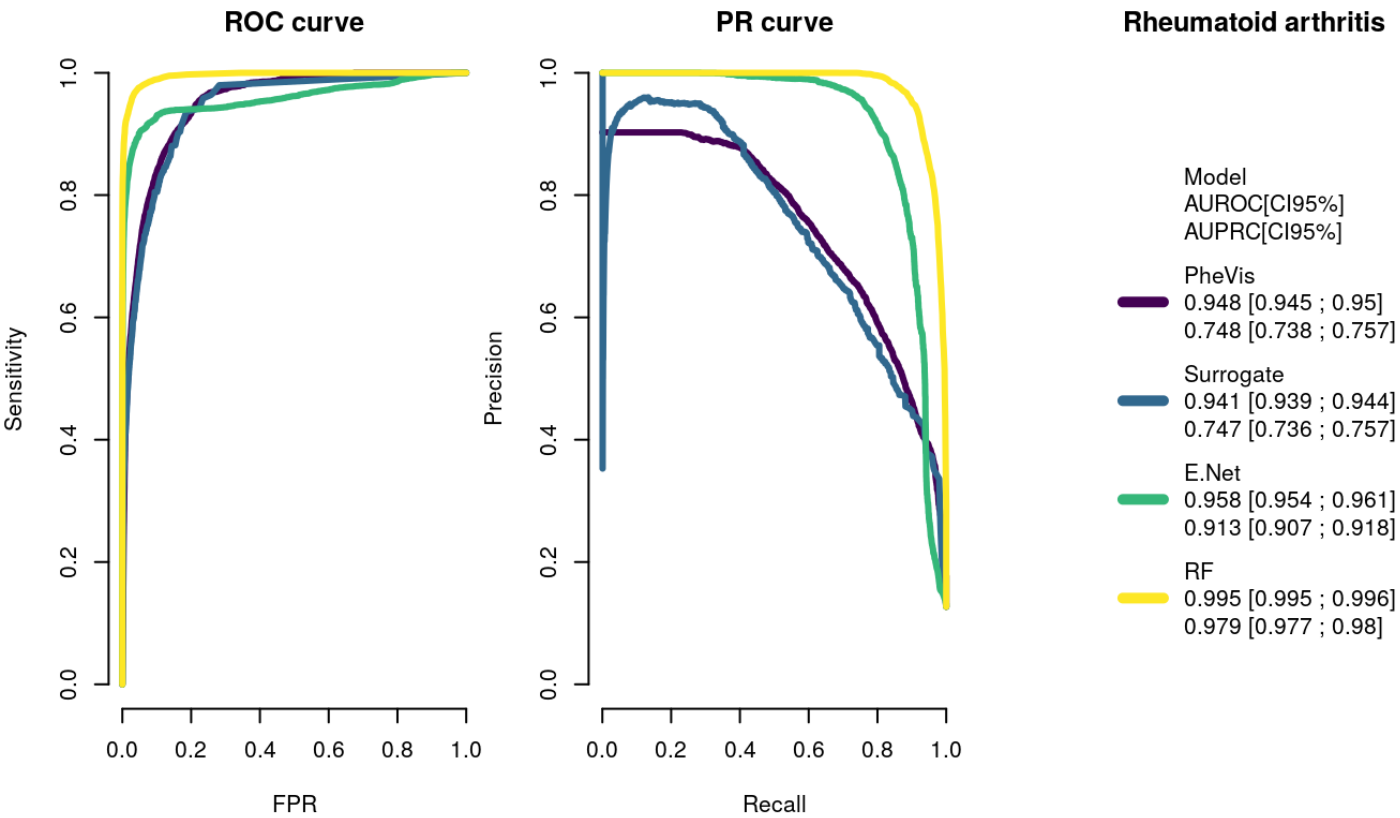
$$\hat{h}: \boldsymbol{X'_{ij}} \rightarrow S_{ij}\ \forall ij\ as\ S_{ij} = \{0,1\}$$
$$PL_{ij} = \hat{h}(\boldsymbol{X'_{ij}})\ \forall ij$$

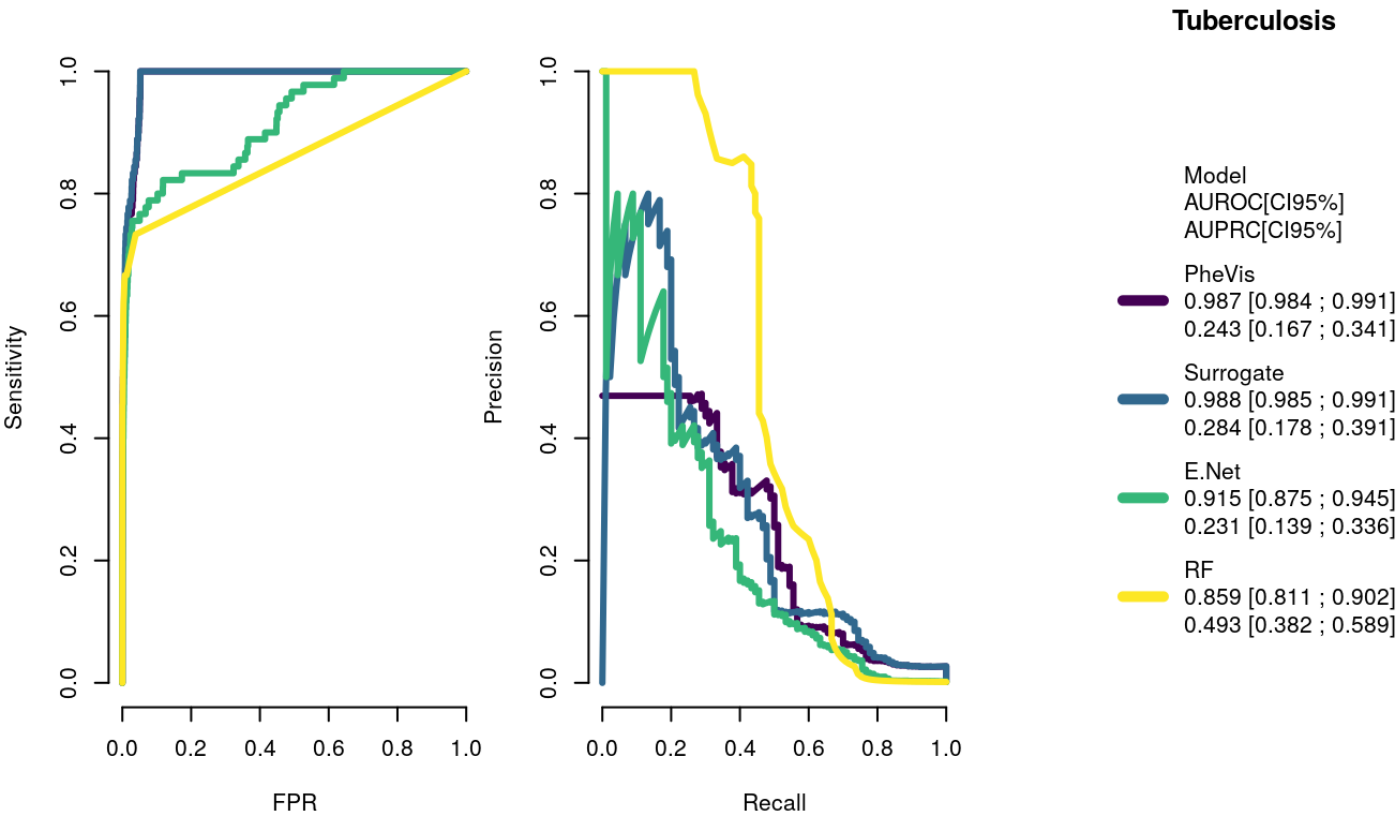
4

Pr: Logistic regression with random intercept for probability prediction.

$$logit\left(P(PL_{ij} = 1)\right) = \boldsymbol{\beta}^T \boldsymbol{X'} + b_{0i}$$
$$b_{0i} \sim N(0, \sigma_0^2)$$


Rheumatoid arthritis
Tuberculosis
patient 1
patient 2
patient 3
patient 4
Probability
1
0
2006
2008
2010
2012
2014
2016
2018
2020
Date
Gold standard
Prediction
Visit


ROC curve
PR curve
Rheumatoid arthritis
Sensitivity
Precision
FPR
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Model
AUROC[CI95%]
AUPRC[CI95%]
PheVis
0.948 [0.945 ; 0.95]
0.748 [0.738 ; 0.757]
Surrogate
0.941 [0.939 ; 0.944]
0.747 [0.736 ; 0.757]
E.Net
0.958 [0.954 ; 0.961]
0.913 [0.907 ; 0.918]
RF
0.995 [0.995 ; 0.996]
0.979 [0.977 ; 0.98]


Tuberculosis
Sensitivity
Precision
FPR
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Model
AUROC[CI95%]
AUPRC[CI95%]
PheVis
0.987 [0.984 ; 0.991]
0.243 [0.167 ; 0.341]
Surrogate
0.988 [0.985 ; 0.991]
0.284 [0.178 ; 0.391]
E.Net
0.915 [0.875 ; 0.945]
0.231 [0.139 ; 0.336]
RF
0.859 [0.811 ; 0.902]
0.493 [0.382 ; 0.589]