

Background

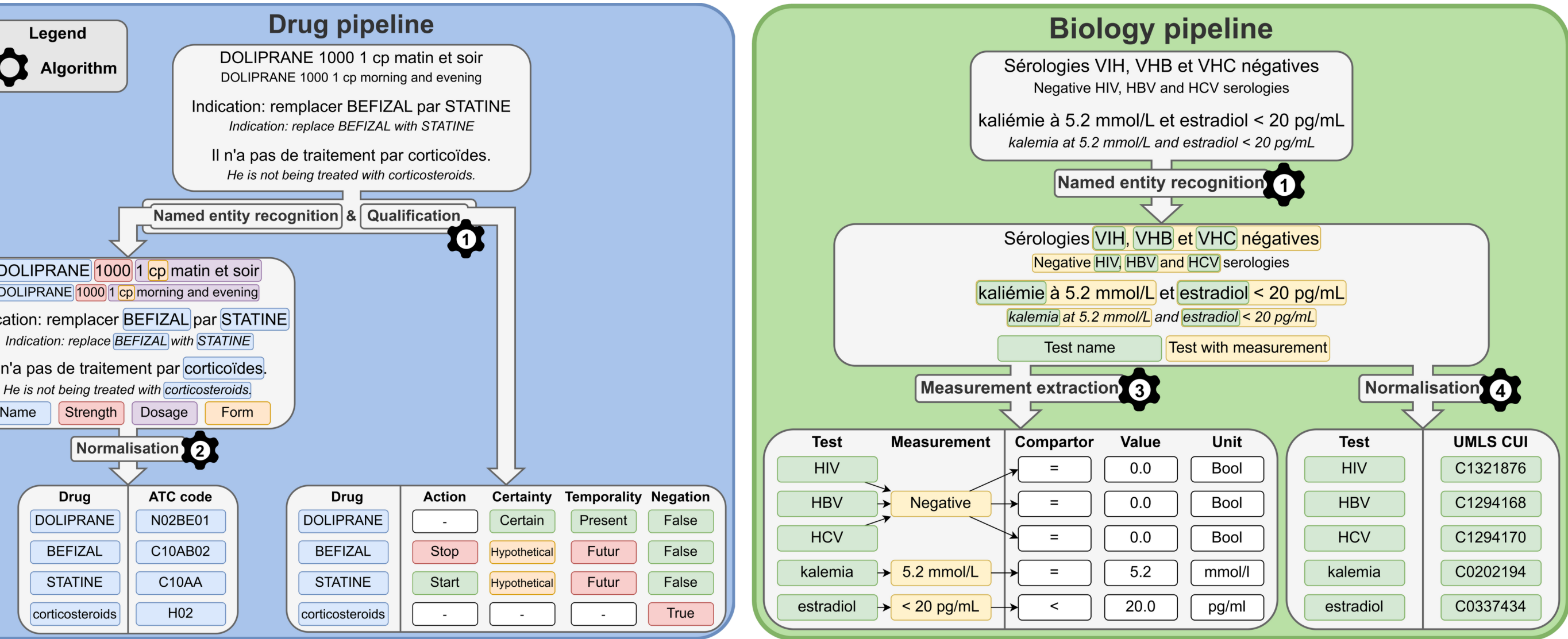
**Patient phenotyping** involves extracting observable characteristics from **electronic health records (EHRs)**, such as biological test results, symptoms, illnesses and treatments, to enable cohort selection, comparison between patients, patient predictive risk assessment, pharmacovigilance analyses, and more.

EHRs contain **structured data** (numerical values and disease codes) and **unstructured data** (clinical texts). However, structured data may be partial, particularly regarding past diagnoses and out-of-hospital results. **Natural Language Processing (NLP)** is crucial for extracting valuable information from clinical texts that is not captured in structured data.

Methods

**DATASET**  
The dataset includes **46 940 patients**: all patients aged over 15 with **systemic lupus erythematosus**, **systemic sclerosis**, **antiphospholipid syndrome** or **Takayasu disease** who had at least one visit to AP-HP hospitals initially over the period 07/01/2017 to 12/31/2020. **103 discharge summaries** have been annotated.

**NATURAL LANGUAGE PROCESSING**  
Raw text is processed by 4 algorithms in total:



Extra results

NAMED ENTITY RECOGNITION

|         |                           | CamemBert-Base |           |        | CamemBert-Bio |           |        | DrBert |           |        | CamemBert-EDS |           |        |       |
|---------|---------------------------|----------------|-----------|--------|---------------|-----------|--------|--------|-----------|--------|---------------|-----------|--------|-------|
|         |                           | N_entity       | Precision | Recall | F1            | Precision | Recall | F1     | Precision | Recall | F1            | Precision | Recall | F1    |
| Bio     | Bio_Test_Name             | 1470           | 86.22     | 86.22  | 86.22         | 89.5      | 88.39  | 88.94  | 86.37     | 86.3   | 86.33         | 89.5      | 91.02  | 90.25 |
|         | Bio_Test_with_measurement | 1258           | 79.05     | 78.67  | 78.86         | 83.55     | 79.54  | 81.5   | 81.39     | 76.46  | 78.85         | 83.15     | 82.04  | 82.59 |
| Drug    | Drug_Dosage               | 306            | 88.89     | 86.96  | 87.91         | 86.81     | 85.87  | 86.34  | 79.48     | 77.17  | 78.31         | 87.05     | 87.68  | 87.36 |
|         | Drug_Form                 | 202            | 82.02     | 85.88  | 83.91         | 83.15     | 90.0   | 86.44  | 82.56     | 83.53  | 83.04         | 86.34     | 92.94  | 89.52 |
| Drug    | Drug_Name                 | 671            | 85.74     | 88.38  | 87.04         | 90.03     | 89.57  | 89.8   | 86.14     | 84.96  | 85.54         | 90.73     | 91.97  | 91.34 |
|         | Drug_Strength             | 141            | 91.73     | 93.85  | 92.78         | 90.51     | 95.38  | 92.88  | 83.7      | 86.92  | 85.28         | 93.94     | 95.38  | 94.66 |
| Overall |                           | 4048           | 84.23     | 84.66  | 84.44         | 87.37     | 86.09  | 86.72  | 84.07     | 82.31  | 83.18         | 87.67     | 88.49  | 88.08 |

NORMALISATION

|           | Model          | CODER ALL            | FUZZY JW             | FUZZY LEV            | SAPBERT ALL          | SAPBERT EDS          |
|-----------|----------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| data type | Dataset        |                      |                      |                      |                      |                      |
| Biology   | Annotated EHRs | 0.726 [0.682, 0.768] | 0.53 [0.422, 0.592]  | 0.548 [0.44, 0.616]  | 0.696 [0.613, 0.736] | 0.704 [0.607, 0.754] |
|           | EMEA           | 0.9 [0.647, 1.0]     | 0.75 [0.499, 0.947]  | 0.65 [0.412, 0.909]  | 0.75 [0.5, 1.0]      | 0.8 [0.5, 0.967]     |
|           | MEDLINE        | 0.741 [0.639, 0.842] | 0.593 [0.481, 0.701] | 0.654 [0.552, 0.773] | 0.679 [0.575, 0.782] | 0.654 [0.539, 0.769] |
| Drug      | Annotated EHRs | 0.788 [0.739, 0.849] | 0.814 [0.763, 0.867] | 0.811 [0.761, 0.869] | 0.773 [0.708, 0.832] | 0.769 [0.708, 0.831] |
|           | EMEA           | 0.922 [0.853, 0.97]  | 0.903 [0.84, 0.955]  | 0.903 [0.843, 0.952] | 0.919 [0.856, 0.972] | 0.918 [0.851, 0.969] |
|           | MEDLINE        | 0.908 [0.873, 0.942] | 0.823 [0.779, 0.866] | 0.82 [0.775, 0.862]  | 0.848 [0.809, 0.889] | 0.87 [0.832, 0.909]  |



# Text mining of electronic health records enhances patient phenotyping.

## 3437 patients with lupus

