LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

A brief Overview of Clinical Texts Automatic De-Identification/Pseudonymization

In: *Legislation and regulations for data spaces: an environment for the development of a European Data Market*

Cyril GROUIN

2024, January 29th

Definitions

- **De-identification or pseudonymization**
  - Explicit identifiers *(I saw Mr Paul Smith on January 29th for left lower back pain)*
    → indirect identifiers *(I saw Mr John Doe on May 23rd for left lower back pain)*
  - Existing list of types of identifiers
    (names, address, dates, record numbers, etc.; cf. HIPAA, 1996)
  - Reversible operation if access to additional data
- **Anonymization**
  - Meystre et al (2010): *“data cannot be linked to identify the patient”*
  - Should not be reversible: more complex, how can we guarantee it?

Clinical Texts
Access Issues

Hospitals

- Medical staff produce data about patients
  - Structured data (databases):
    → easy to process
  - Unstructured data (texts):
    → needs for ad hoc NLP tools
- Clinical records: sensitive data
  - Data protection needed within the hospital
  - Pseudonymization/anonymization for use out of hospital

Clinical Texts
Access Issues

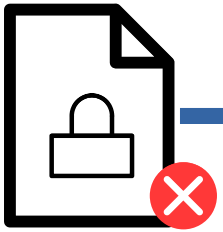
Hospitals

- Medical staff produce data about patients
  - Structured data (databases):
    → easy to process
  - Unstructured data (texts):
    → needs for ad hoc NLP tools
- Clinical records: sensitive data
  - Data protection needed within the hospital
  - Pseudonymization/anonymization for use out of hospital

Academics

- Needs real data
  - To study clinical language properties
  - To produce/evaluate NLP tools
- Can hardly access clinical records (GDPR)
- Alternative solutions are similar but not real (distinct language properties):
  - Clinical cases
  - Generation of clinical records

Clinical Texts Pseudonymization Tool



Clinical record with sensitive data

How to process?

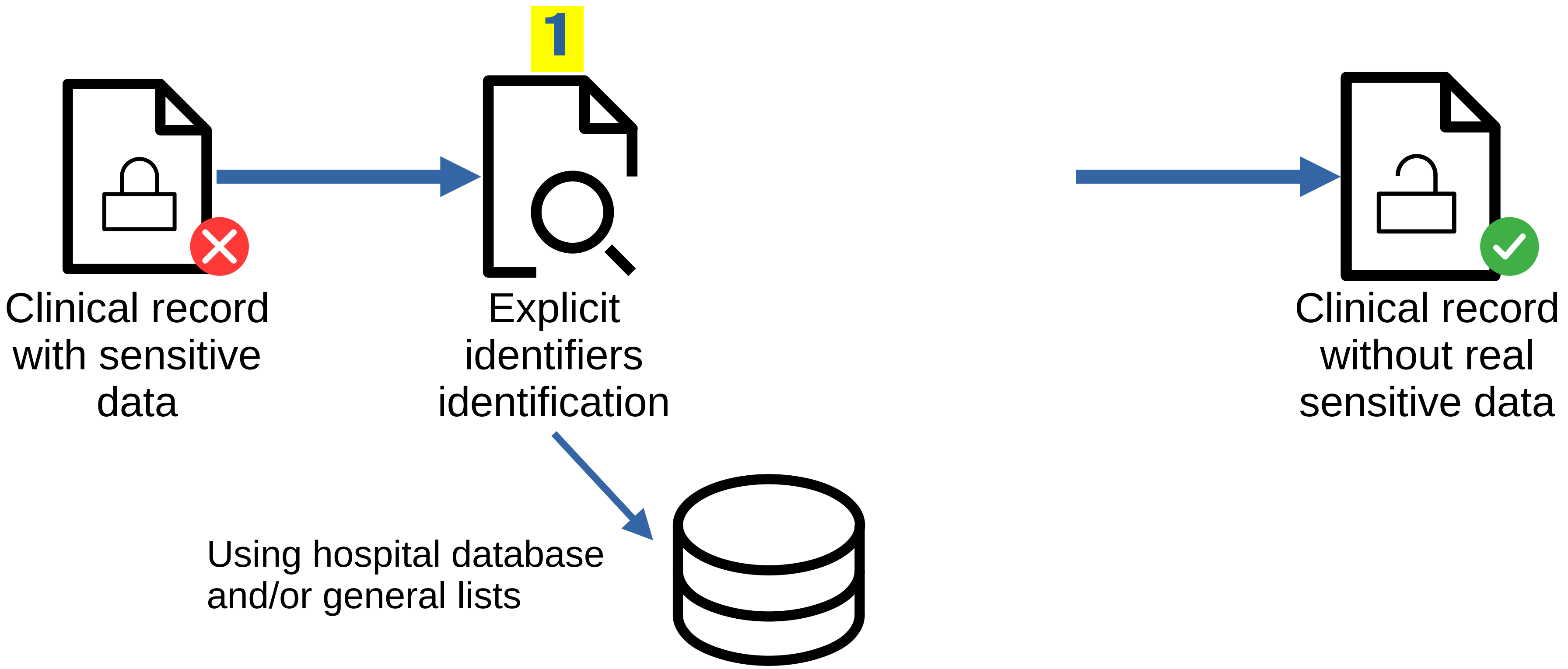
Clinical record without real sensitive data

Clinical Texts Pseudonymization Tool

Two main stages:

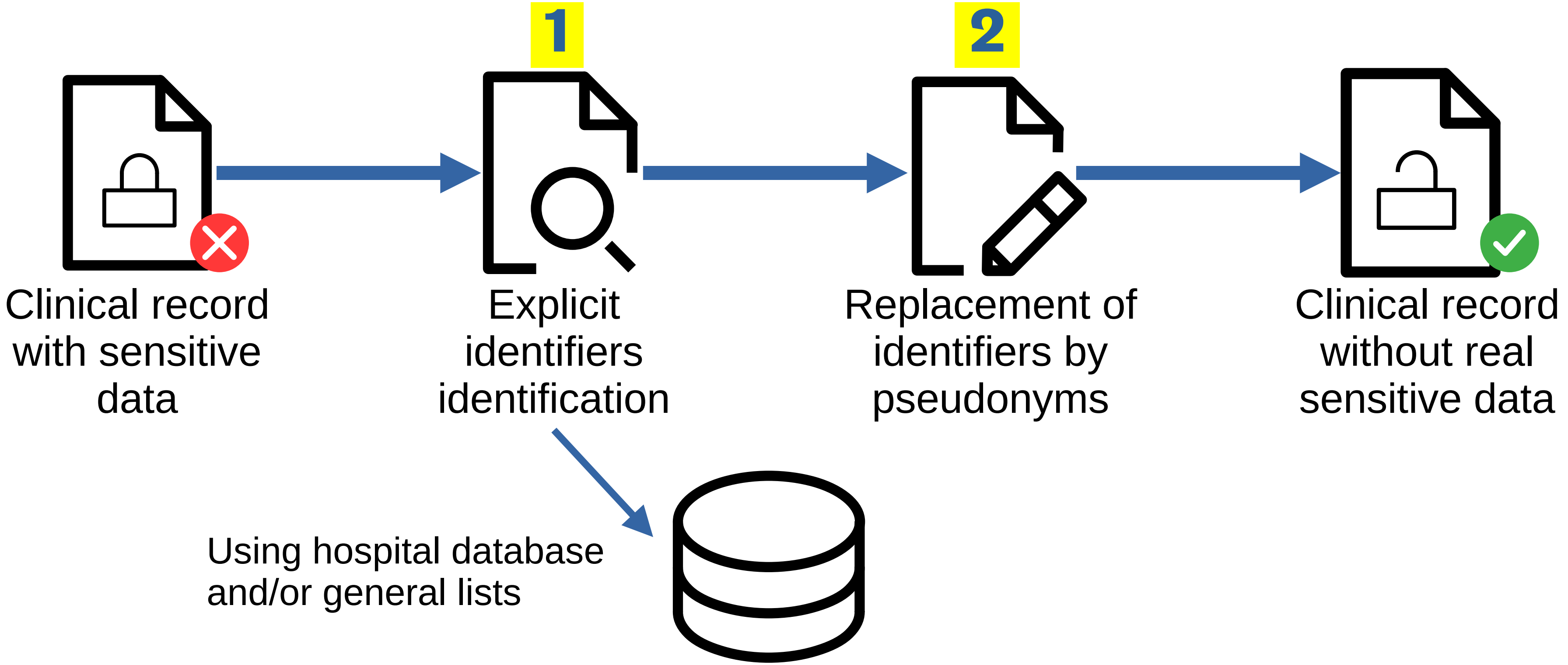
Clinical Texts Pseudonymization Tool

Two main stages:



Clinical Texts Pseudonymization Tool

Two main stages:



Clinical Texts
1 – Identifying the identifiers

Data format	Types of identifiers	Complexity level	Useful approach
Numerical	Date, Phone number, etc.	Low • Regular format • Low diversity • Rare ambiguity	• Regular expressions • Statistical approaches

Ok human check needed

Clinical Texts
1 – Identifying the identifiers

Data format	Types of identifiers	Complexity level	Useful approach
Textual	Names, City	Medium • Exhaustiveness • Possible ambiguity • Out-of-Vocabulary	• Lists / Hospital database • Context-based regular expressions • Statistical approaches

Ok human check needed

Clinical Texts
1 – Identifying the identifiers

Data format	Types of identifiers	Complexity level	Useful approach
Textual			
	Address, Facility	❌ High • Huge diversity (format, length)	• Context-based regular expressions (still complex) ❌ + • Statistical approaches ⚠️

✅ Ok ⚠️ human check needed

Clinical Texts
1 – Identifying the identifiers

Data format	Types of identifiers	Complexity level	Useful approach
Numerical	Date, Phone number, etc.	Low • Regular format • Low diversity • Rare ambiguity	• Regular expressions • Statistical approaches
Textual	Names, City	Medium • Exhaustiveness • Possible ambiguity • Out-of-Vocabulary	• Lists / Hospital database • Context-based regular expressions • Statistical approaches
	Address, Facility	High • Huge diversity (format, length)	• Context-based regular expressions (still complex) + • Statistical approaches

Ok human check needed

Clinical Texts
2 – Pseudonymizing the content

Type of identifier	Process	Clinical usefulness	Points of vigilance
Date, Age	Date shiffting	Intervals of dates are kept (clinical value)	• New dates must be relevant from a clinical point of view (infant age w/ adult disease; Covid-19 in 2016) • Adding a slight random shiffting process in intervals of dates (differential privacy)
Phone, Zip, Medical record number	Random draw	No	
Names, Address, City, Hospital names	Random draw	Demographic data (importance of a city, known impact of the environment on health)	• Retain original data distribution for further statistical-based NLP

Evaluation
How to evaluate de-identified outputs?

Classical way in NLP:

- Are sensitive data correctly identified?
  - True positive, False positive, False negative
    → Recall (Sensitivity), Precision, F1-score
  - Frontiers: *I saw Mr [John] Doe*
    → Incorrect span
  - Labels: *Parking Office Customer service [32330]*
    → Zip (incorrect) vs. Telephone (5-digit extension of a main phone number)

Evaluation
How to evaluate de-identified outputs?

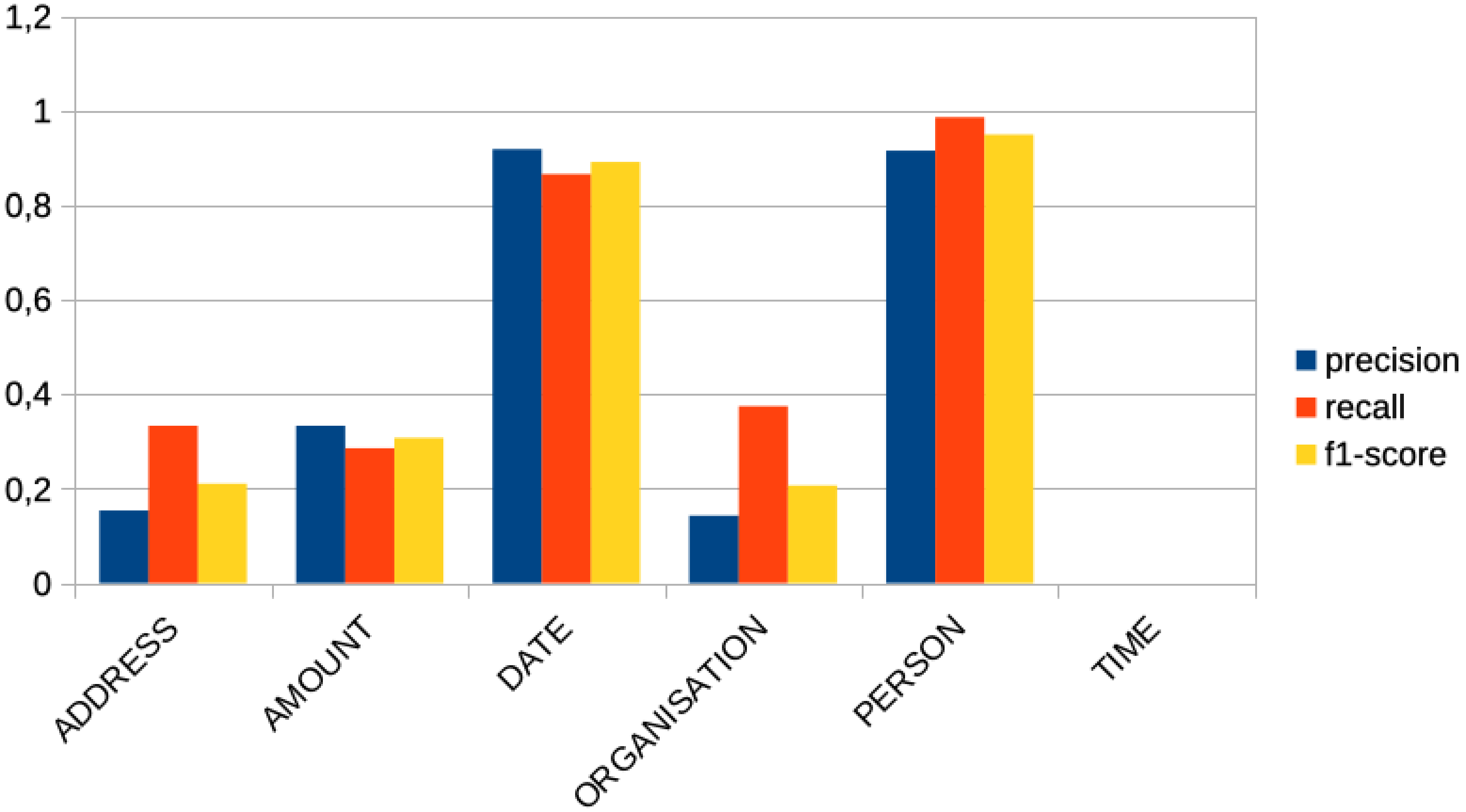
Classical way in NLP:

- Are sensitive data correctly identified?
  - True positive, False positive, False negative
    → Recall (Sensitivity), Precision, F1-score
  - Frontiers: *I saw Mr [John] Doe*
    → Incorrect span
  - Labels: *Parking Office Customer service [32330]*
    → Zip (incorrect) vs. Telephone (5-digit extension of a main phone number)

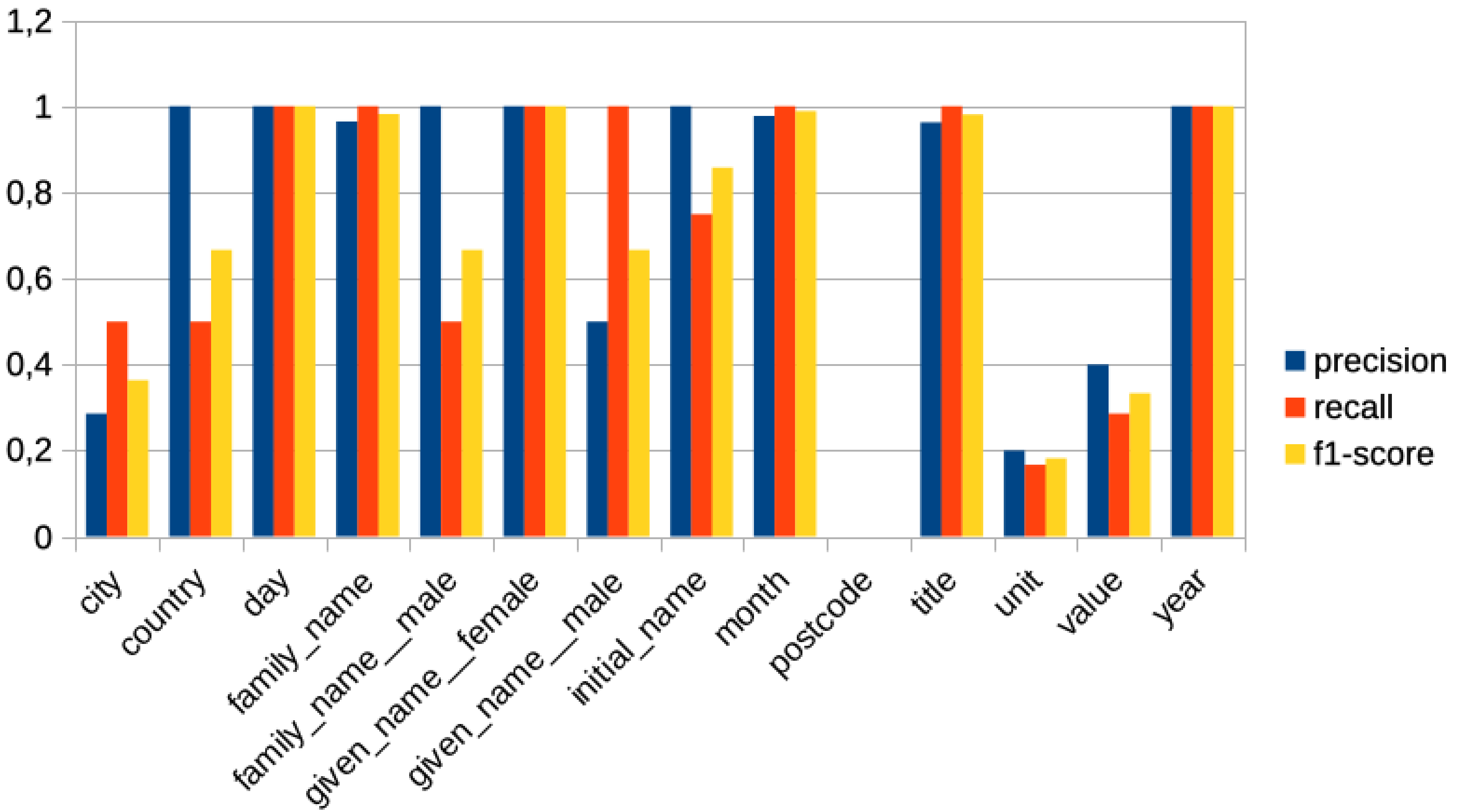
Clinical-oriented evaluation:

- Are pseudonymized data still clinically useful?
  (e.g., if all patient and doctor names are replaced by *John Doe*
  → loss of information)
- If one sensitive data is not pseudonymized (e.g., patient's last name),
  is the whole clinical record still pseudonymized?

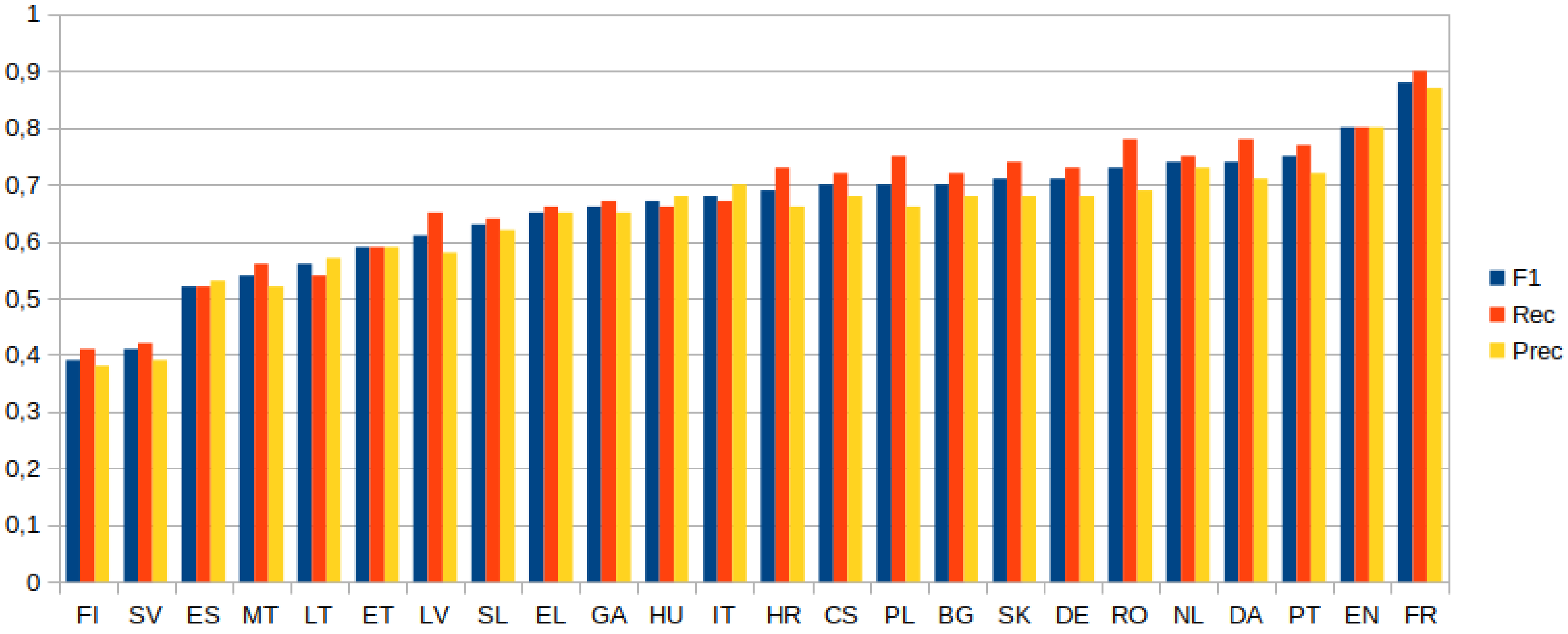
Evaluation
French Clinical Cases – Main Labels



Evaluation
French Clinical Cases – Detailed Labels



Evaluation
Global results per language



French clinical cases translated into 26 langages, then de-identified

Conclusions

- Data are stored in hospital, researchers are in academic labs ❌
  - But a few medical doctors succeed to use AI tools and models on their data (especially medical staff with NLP PhD) ✅
- Why de-identifying clinical data?
  - Who will access the data? Which objective?
  - Does the final user need to keep a consistent replacement of explicit identifiers in all documents from a given patient?
- How de-identifying sensitive data?
  - Depending on the type of sensitive data, using transformers models (that have a strong impact on the environment) is not always the best solution ❌