Leveraging Information Redundancy of Real-World Data Through Distant Supervision

Presentation: Ariel Cohen

Innovation and Data Unit, IT Department, AP-HP

Ariel Cohen – Innovation and Data Unit, IT Department, AP-HP
Alexandrine Lanson – LIMICS – Sorbonne Université
Emmanuelle Kempf - Department of Medical Oncology, Henri Mondor and Albert Chenevier Teaching Hospital, AP-HP
Xavier Tannier - LIMICS – Sorbonne Université

Introduction

NLP applied to **Health Care** notes

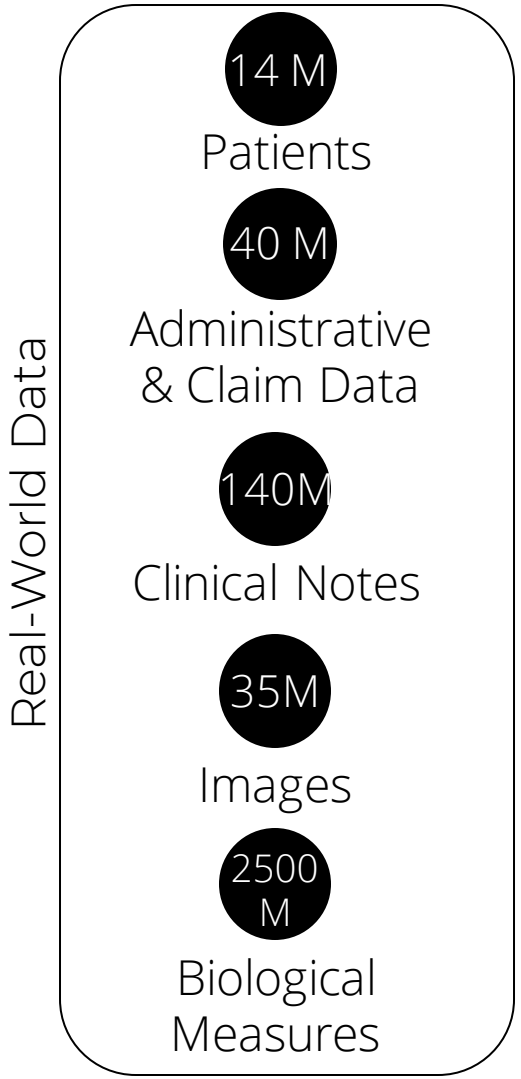
✓ Satisfactory results

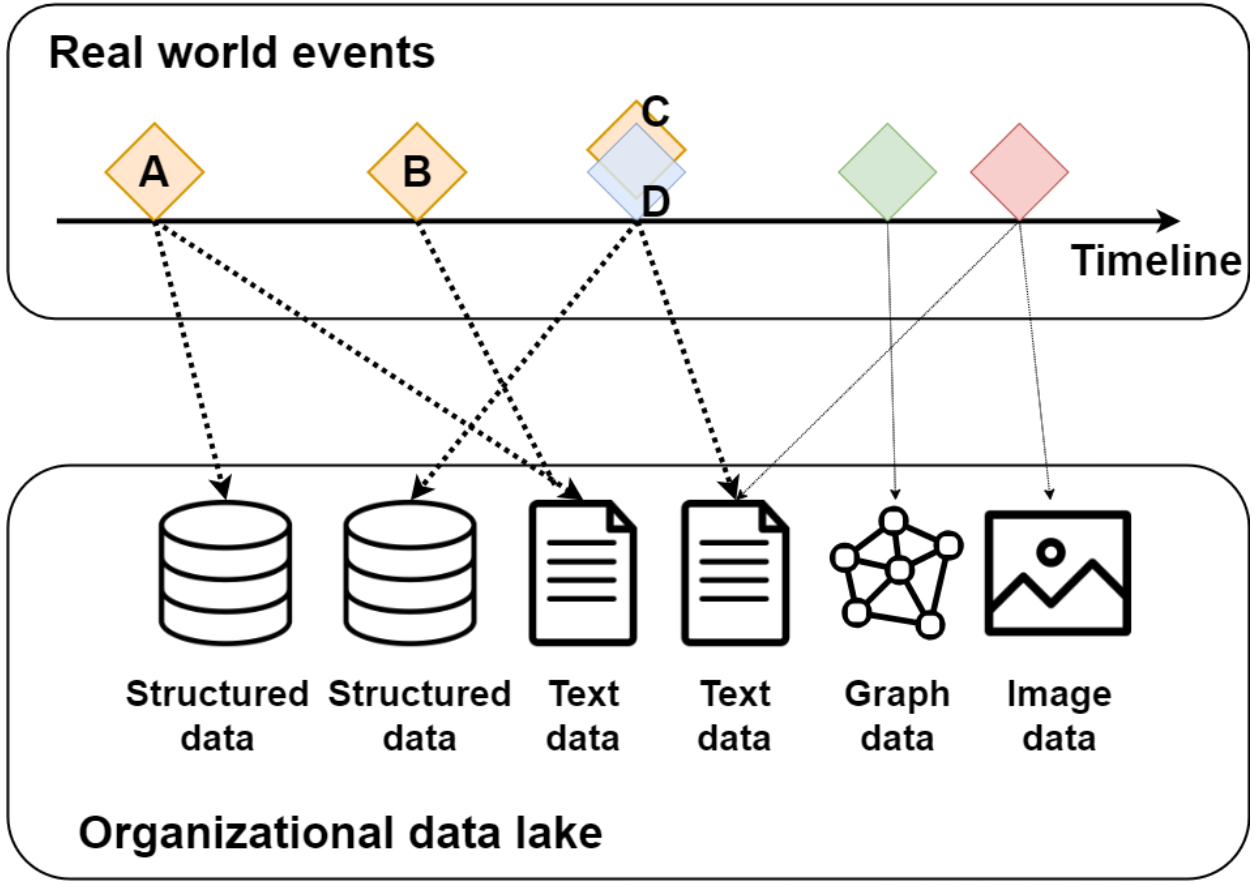
✗ Supervised approaches dependent on many **annotated records**

-  Performed by domain experts
  - Scarce and expensive resource
-  Subject to RGPD
  - Not possible to share

→ Bottleneck for research

Paris University Hospitals CDW & Information Redundancy





Distant Supervision Approach

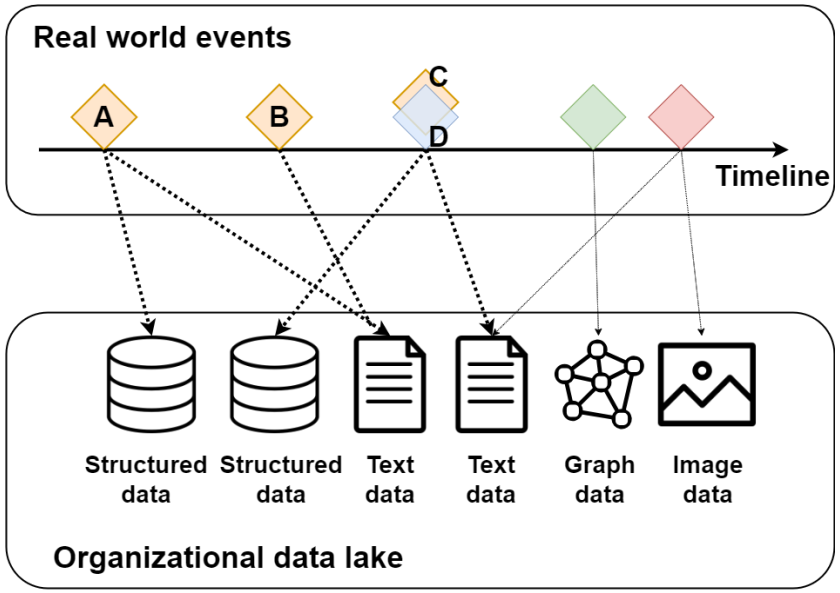
- Use multiple data sources to build annotated datasets
- Faster than what can be produced by manual annotation
- Produce a **silver standard** dataset with **Noisy Labels**
- Deep neural networks (DNNs) are susceptible to **overfit on noisy labels**
- Several efforts and **methods** have been developed to be able **to learn from noisy labels**

Objectives & Outline

- We propose a new **text labeling method** that **leverages temporal information redundancy** from external sources without using any other text information apart from dates themselves for **the task of event extraction and classification**
- **Apply and test** this method in order to reconstruct key events of cancer patients' pathway using clinical notes of a **cohort of 380,000 patients**. We train a classifier using different methods for **noise label management**, **validate the end-to-end approach** and compare it with a **baseline** classifier using an **expert annotated corpus**
- Use this model for **inference** where information is not redundant or lacking

Methodology

- We consider a **task** of **event extraction and classification**, from an organizational data lake **containing text and other sources** of information about real-world events.
- Under the assumption that yellow type events are mentioned using a similar language distribution, we **will leverage the text around the mention of the first event to train a model which allows us to identify the second yellow event from text**.
- We use the **time dimension as a pivot** to align complementary information present in different data sources



Programmatic Annotation : **Positive Examples**

- General method for date alignment between an external source of **structured data** including event dates and **a paired text corpus** (e.g., linked by a person/client/user identifier).

Positive Examples

Algorithm

1. Date recognition (NER) & Normalization
2. Select dates mentioned once in the document
3. For each person join the dates from structured data with the ones of step 1 >> **aligned dates**

Person ID	Date	Medical Procedure
1	18/8/2015	Surgery
1	20/07/2019	Biopsy
2	3/10/2022	Biopsy

A male patient born on 5/1/1991 with a tumour in the upper pole of the right testicle found by chance during self-exploration, for which reason he attended a urology consultation on 5/10/2020 where a physical examination was performed,
[...] The patient was biopsied on 20/7/2019 at the hospital [...]

Linked by Person ID

Programmatic Annotation : **Negative Examples**

2 methods for the selection **of alternative examples**

1. Random Selection (RS): A naive random selection among all non-matched dates in the previous step;
2. Proximity Selection (PS): A random selection weighted by a score of distance in text with respect to the matched dates.

A male patient born on 4/1/1991 with a tumour in the upper pole of the right testicle found by chance during self-exploration, for which reason he attended a urology consultation on 5/10/2020 where a physical examination was performed,
[...] The patient was biopsied on 20/7/2019 at the hospital [...]

Dataset Creation & Illustrative example

For each labelled date (positive and negative) we extract a **textual context surrounding the date** (snippet) and we **mask** the date to be classified

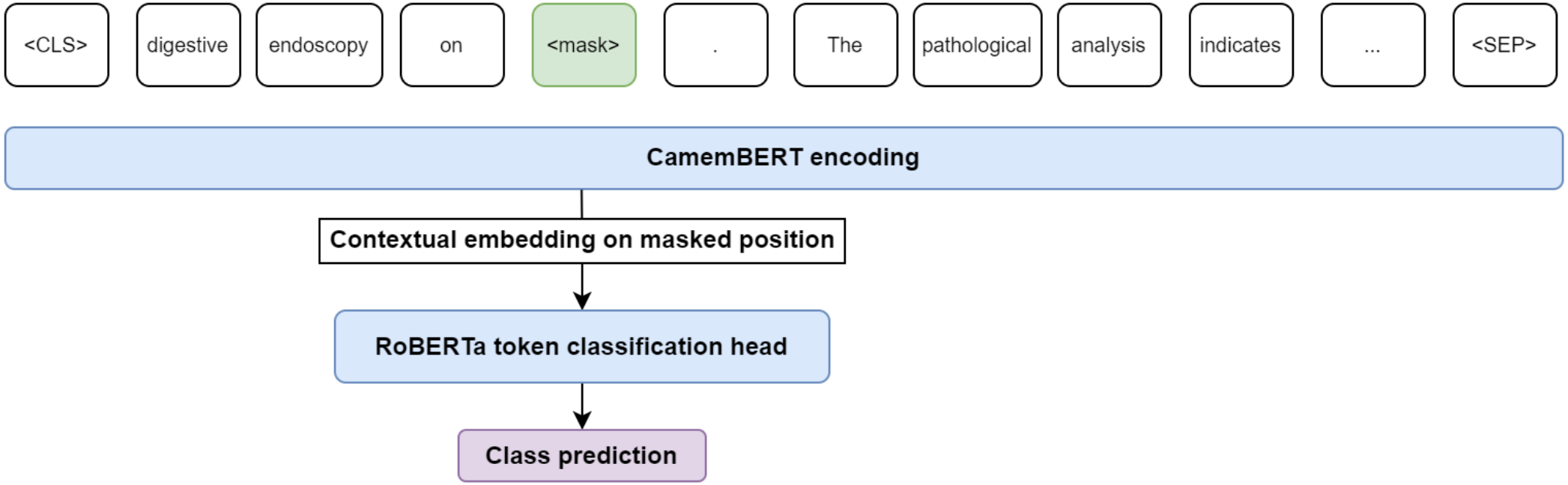
>> **Silver dataset**

>> Use of **Noise Management** techniques to **learn** from it

ID	Text	Silver label	Gold label
1	The patient was biopsied on <masked date> at the hospital	Biopsy	Biopsy
2	The patient showed the results of the biopsy done outside the hospital on <masked date> [..]	Other	Biopsy
3	digestive endoscopy on <masked date>. The pathological analysis indicates [..] On 17.08.2019 the patient underwent abdominal surgery	Biopsy	Biopsy
4	digestive endoscopy on February 15th, 2019. The pathological analysis indicates [..] On <masked date> the patient underwent abdominal surgery	Surgery	Surgery
5	Mr. Smith came to consultation on <masked date>	Other	Other
6	Mrs. Dupont came to consultation on <masked date>	Biopsy	Other

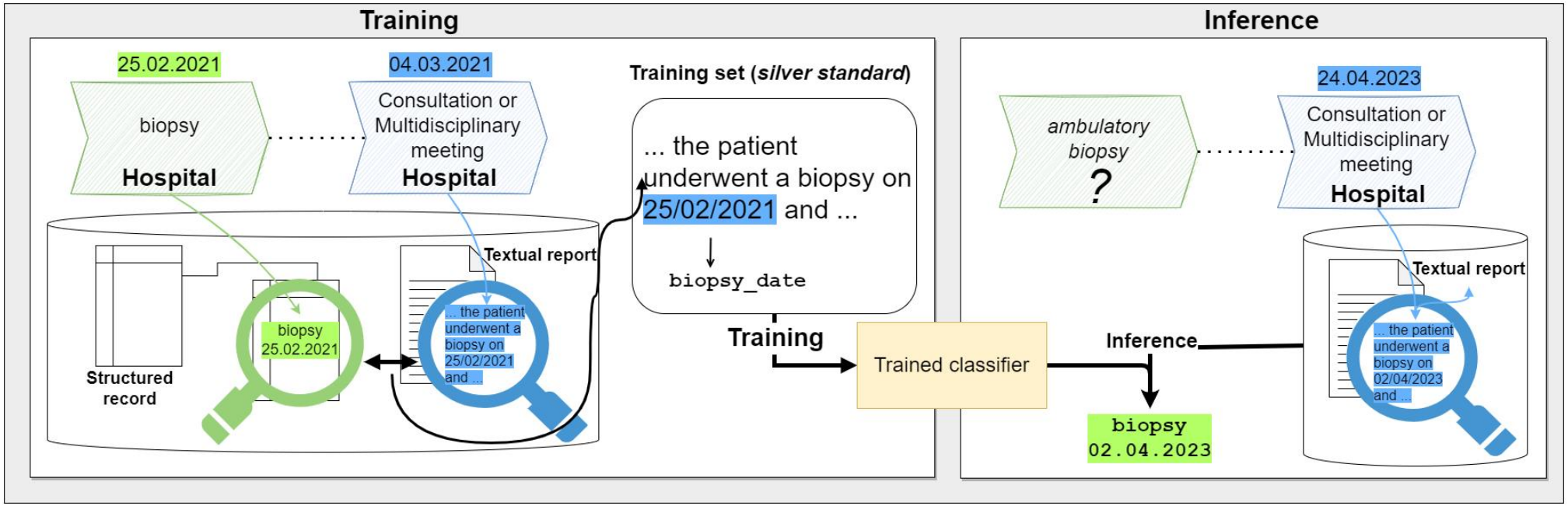
Modelling

- CamemBERT pre-trained encoder
- RoBERTa Token Classification Head



Application use case

- The CDW of the Greater Paris University Hospitals (Assistance Publique-Hôpitaux de Paris — AP-HP) contains the EHR of 380,000 patients with cancer and around 35,000 new cancer cases each year
- We address the problem of cancer patient reconstruction and focus on diagnosis date (biopsy) because of its importance in oncology research (e.g., implication in treatment effectiveness studies, survival analysis, epidemiology, etc.).



Experiment Setting

- **Test** and **Development** corpus (101 & 60 documents respectively).
- This corresponds to **1,474 date entities** annotated in a binary way for the **test set** and **680** for **the development set** (**4.5%** and **10%** correspond to biopsies respectively).
- Apply the **programmatic annotation method** on 4 different settings (data sources and negative examples selection)
- For each produced dataset we test different training configurations:
  - Regular training with CE loss
  - NCE-RCE loss
  - O2U strategy with CE loss
  - O2U strategy with NCE-RCE loss
  - LRT strategy
- Multiple training with different number of examples of the **development** set to measure **performance as a function of the size** of the **expert annotated** data
- Compare the performance within patients with and without a biopsy done at hospital

Results

- 2 training datasets of 10,850 samples (50% biopsies) and 2 of 16,200 samples (33% biopsies).

Method	PR-RS	PR-PS	PR-SP-RS	PR-SP-PS
CE	0.54 (0.53 - 0.54)	0.49 (0.44 - 0.54)	0.46 (0.45 - 0.48)	0.44 (0.42 - 0.45)
NCE-RCE	0.55 (0.53 - 0.60)	0.58 (0.55 - 0.60)	0.60 (0.58 - 0.70)	0.69 (0.40 - 0.74)
O2U - NCE-RCE	0.51 (0.46 - 0.58)	0.56 (0.50 - 0.61)	0.68 (0.67 - 0.71)	0.64 (0.63 - 0.71)
O2U - CE	0.58 (0.50 - 0.59)	0.56 (0.54 - 0.59)	0.67 (0.60 - 0.72)	0.68 (0.63 - 0.71)
LRT	0.56 (0.47 - 0.62)	0.58 (0.56 - 0.62)	0.65 (0.57 - 0.70)	**0.70** (0.65 - 0.70)

Table 2: Median F1-score ([min-max] over 5 iterations on the test set) comparison between methods using different training datasets. These are: i. PR-RS: Pathology Reports and Random Selection; ii. PR-PS: Pathology Reports and Proximity Selection; iii. PR-SP-RS: Pathology Reports, Surgery Procedures and Random Selection; iv. PR-SP-PS: Pathology Reports, Surgery Procedures and Proximity Selection.

Class	Precision	Recall	Support
Biopsy	0.91	0.62	34
Other	0.96	0.99	280

Table 3: Performance of the programmatic annotation for the PR-RS dataset, evaluated on development set.

Results: a baseline with an expert annotated dataset

- Model performance increases in function of the number of examples used for training using an expert annotated dataset
- **The distant supervised approaches perform comparably to the results obtained with 450 clean labeled entities**
- Performance increase is slower when using more than 20 documents.

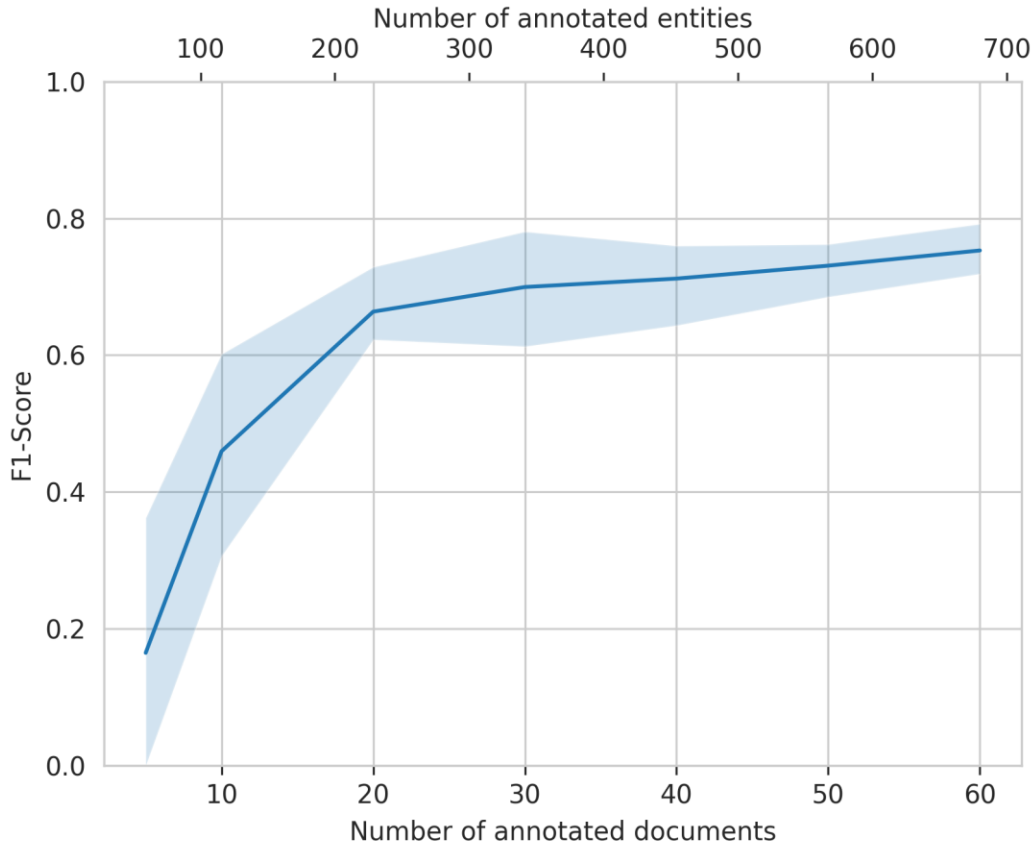


Figure 3: Median F1-score ([min-max] over 5 iterations) in function of number of training examples using an expert labeled dataset.

Results: bias

- Model performance for two patient groups: those who underwent at least one hospital biopsy, and the others.
- Higher recall rates among the first group

	w/ biopsy		w/o biopsy	
Method	**Prec.**	**Rec.**	**Prec.**	**Rec.**
CE	0.38	0.90	0.27	0.53
NCE-RCE	0.61	0.90	0.59	0.69
O2U - NCE-RCE	0.54	0.93	0.57	0.69
O2U - CE	0.56	0.93	0.60	0.69
LRT	0.64	0.90	0.59	0.69
EL dataset - CE	0.69	0.93	0.62	0.75

Table 4: Median precision and recall (over 5 iterations) for patients with and without a biopsy done at the hospital, evaluated on test set. Model trained on PR-SP-PS and EL datasets.

Discussion & Conclusions

- Our **domain agnostic** method helps to **minimize** the necessary **annotation effort** in a context when **experts' available time is scarce**, such as healthcare; however, we reduce expert annotation effort **in exchange** for **an industry knowledge** investment.
- All models, including the expert labeled one, **underperform** when evaluated on **patients without biopsy procedures** done at the hospital.
- The **absence of any NER** procedure for the concepts to identify **allows** us to learn **complex patterns** from text
- The method is **language- and vocabulary-independent**, and therefore directly applicable to other CDWs, to retrieve other important dates or in another dataset that includes event dates and a paired text corpus.
- **Successful results comparable** with models fitted on an **expert annotated corpus**
- The benefit and need of using **noise management** methods is also demonstrated. We **improve performance up to 59%**, all approaches are distinctly better than a classic training using the CE loss
- It has been experimentally shown that the use of **multiple sources** combined together reach **better results**
- Other techniques of **noise management** and **semi-supervised** approaches, as well as a better **bias understanding**, are still to be explored.

Questions?

Perspectives?

Suggestions?

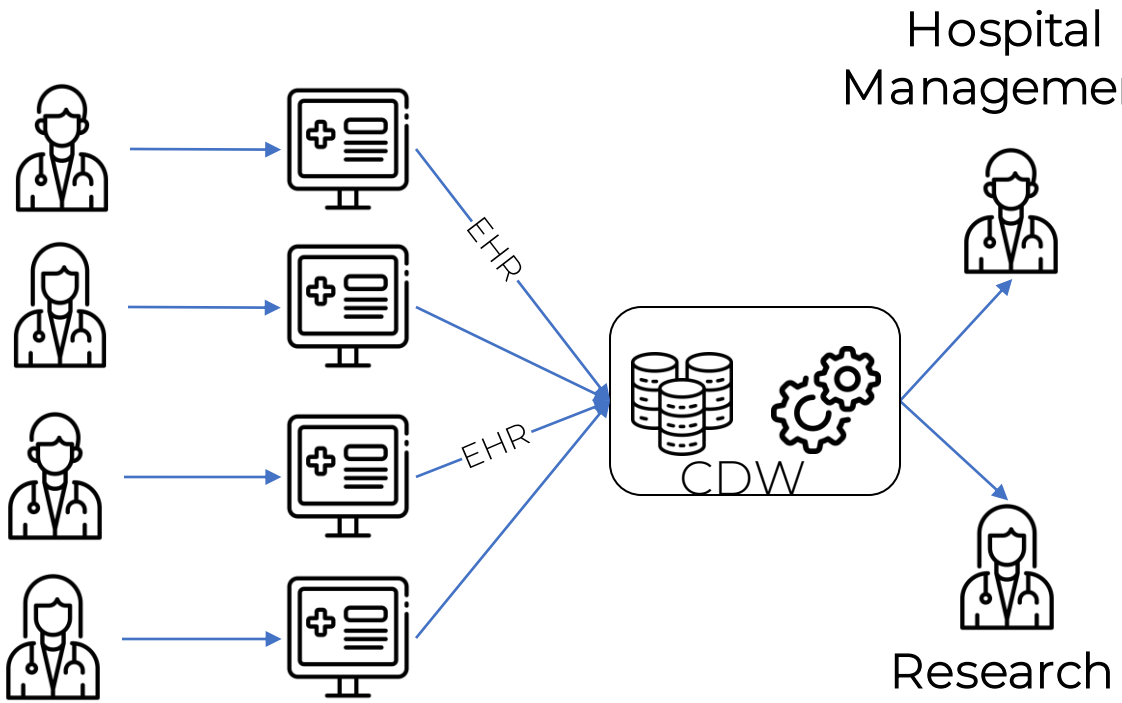
Appendix

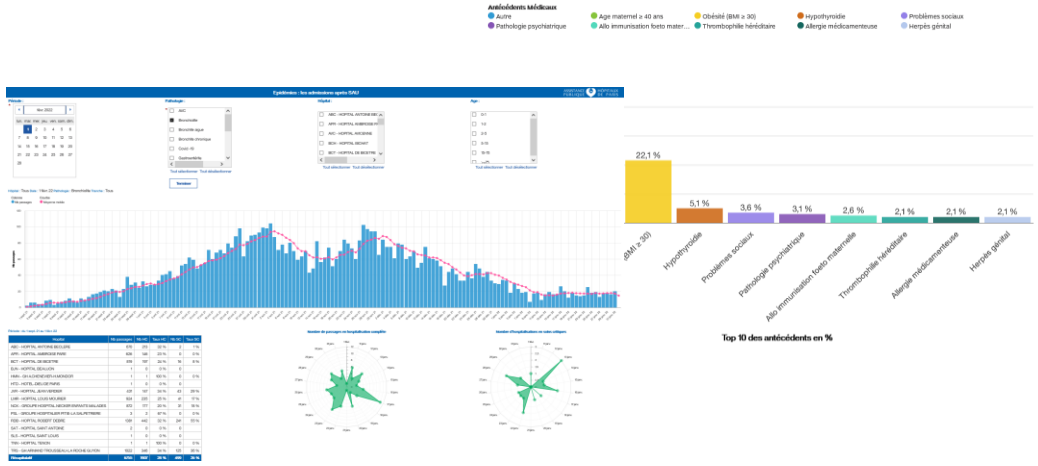
Noise Management

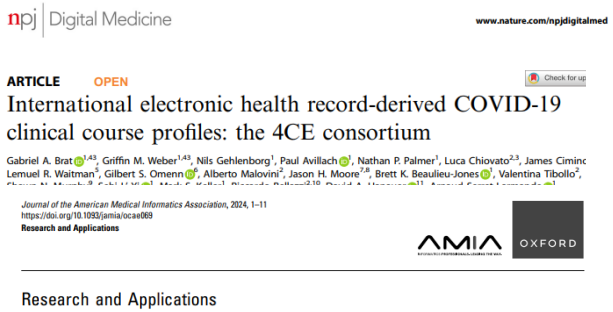
1. Robust Loss Function: **Normalized Cross Entropy - Reverse Cross Entropy** (NCE-RCE) loss, an Active-Passive loss introduced by Ma et al. (2020), with theoretical robustness properties
2. Sample Selection: **O2U-Net** approach (Huang et al.,2019). Multiple rounds changing between **O**verfitting and **U**nderfitting using a cyclical learning rate. Sample losses are ranked and the top k% is then removed from the dataset. Then a final step of training is done on cleaned data.
3. Label Refurbishment: a classifier trained on noisy data has low confidence in the label of a sample, that label is likely to be false Zheng et al. (2020). The method applies a Likelihood Ratio Test (LRT) on noisy classifier predictions to check label purity, the likelihood ratio is compared with a predetermined threshold δ; then, it corrects wrong labels for future training

Clinical Data Warehouse (CDW)

CDW






npj | Digital Medicine

www.nature.com/npjdigitalmed

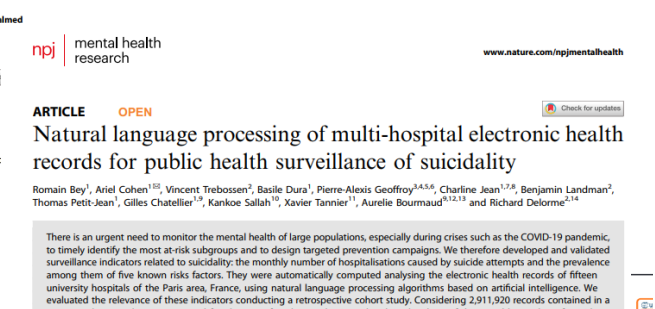
ARTICLE OPEN

International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium

Journal of the American Medical Informatics Association, 2024, 1–11
https://doi.org/10.1093/jamia/ocae069
Research and Applications

Research and Applications

Collaborative and privacy-enhancing workflows on a clinical data warehouse: an example developing natural language processing pipelines to detect medical conditions


npj | mental health research

www.nature.com/npjmentalhealth

ARTICLE OPEN

Natural language processing of multi-hospital electronic health records for public health surveillance of suicidality

There is an urgent need to monitor the mental health of large populations, especially during crises such as the COVID-19 pandemic, to timely identify the most at-risk subgroups and to design targeted prevention campaigns. We therefore developed and validated surveillance indicators related to suicidality: the monthly number of hospitalisations caused by suicide attempts and the prevalence among them of five known risks factors. They were automatically computed analysing the electronic health records of fifteen university hospitals of the Paris area, France, using natural language processing algorithms based on artificial intelligence. We evaluated the relevance of these indicators conducting a retrospective cohort study. Considering 2,911,920 records contained in a common data warehouse, we tested for changes after the pandemic outbreak in the slope of the monthly number of suicide

Cancer Epidemiology

No changes in clinical presentation, treatment strategies and survival of pancreatic cancer cases during the SARS-COV-2 outbreak: A retrospective multicenter cohort study on real-world data

*EHR : Electronic Health Records

Other conclusions

- We confirm that it is possible to **leverage information redundancy** of an **organizational data lake** to **build a programmatically annotated corpus** and train ML models, **minimizing the required expert time for the annotation task**.
- **Domain agnostic** approach, interesting in settings with scarce experts' available time, huge amounts of data and an industry knowledge
- Structuring information from text **impacts downstream applications** as patient recruitment for clinical trials, treatment effectiveness studies, survival analysis or epidemiology studies
- Possible to **share** a method without sharing model weights or datasets