# Differentially Private De-Identification of Clinical Textual Documents: application to ICD10 codes association

*Yakini* Tchouka[1], *Jean-François* Couchot[1], *David* Laiymani[1]
*Philippe* Selles[2], *and Azzedine* Rahmani[2]

[1]Université de Franche-Comté, CNRS, France

[2]Hôpital Nord Franche-Comté, Trevenans, FRANCE

# Plan

# Plan

# Legal context of de-identifying clinical textual document

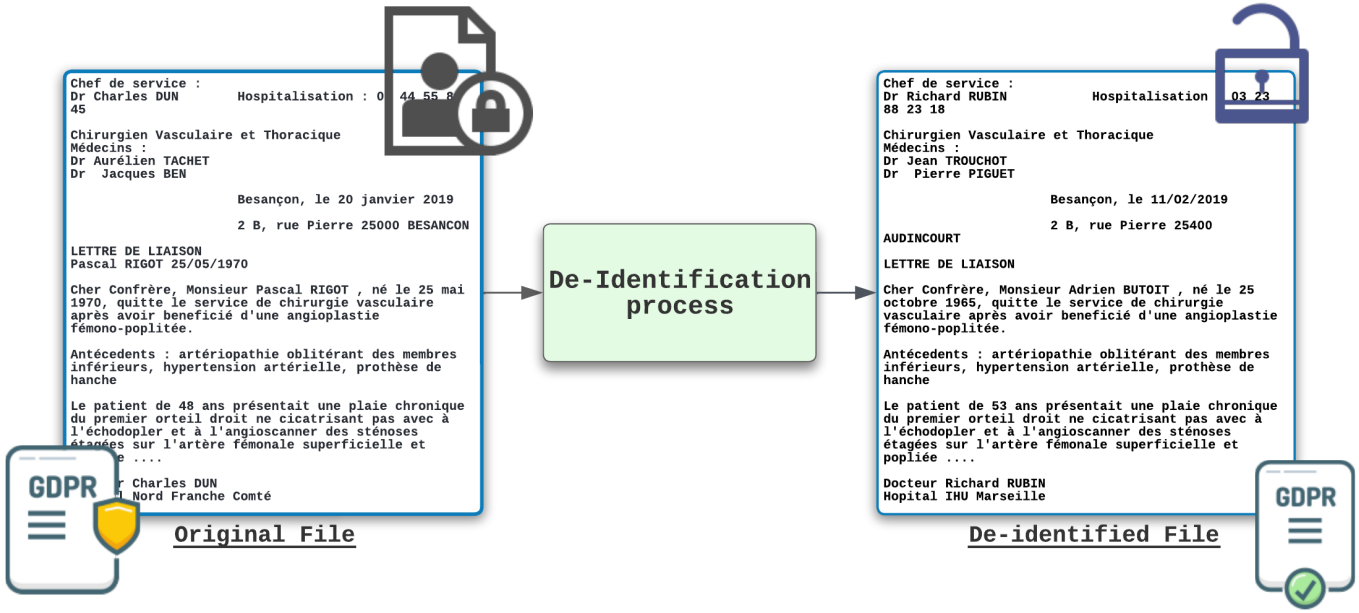
## Considered data type

- Unstructured data: clinical textual documents with name, age, location . . .
  - Natural Language Processing (NLP) task
- Neither ~~images~~ nor ~~tabular data~~

## Legal Requirement

- Make medical data accessible to researchers whilst preserving patients privacy
- Legal requirement imposed by legislation: before sharing
  - GDPR: Delete any data that could identify an individual $\rightsquigarrow$ de-identification
  - HIPAA: List of 18 attributes to remove from a medical document $\rightsquigarrow$ more explicit

# De-identifying: global overview




## Researchers with de-identified data can

- provide models for other medical tasks (clinicalBERT[1], a BERT[2] particularisation)
- Can apply further NLP Task: text summarizing, or here multi-label classification task (ICD-10 codes association)

[1] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

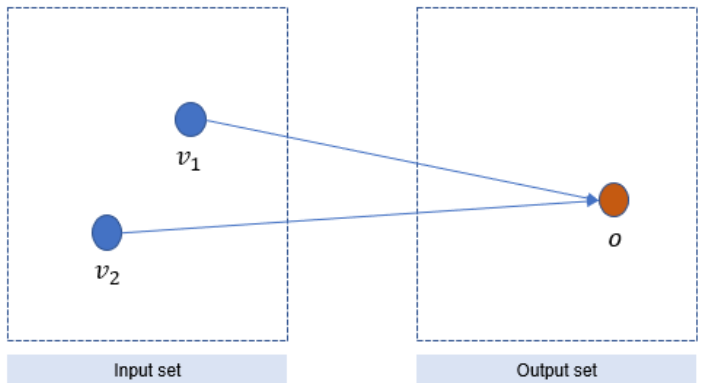
# De-identifying: with Differential Privacy

## Motivation for $\epsilon$ differential privacy [3]

- Method proven on the Sanitization mechanism: robust whatever the data (and therefore the attacks)
- Probabilistic approach valid in the worst-case situation

## Definition ($\epsilon$-local differential confidentiality for a mechanism $\mathcal{M}$)

$\mathcal{M}$ verifies the $\epsilon$-LDP if [4]
$\forall v_1, v_2 \in \text{Domain}(\mathcal{M}), \forall o \in \text{Image}(\mathcal{M})$,
$\Pr[\mathcal{M}(v_1) = o] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(v_2) = o]$.




[3] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3 (pp. 265-284). Springer Berlin Heidelberg.

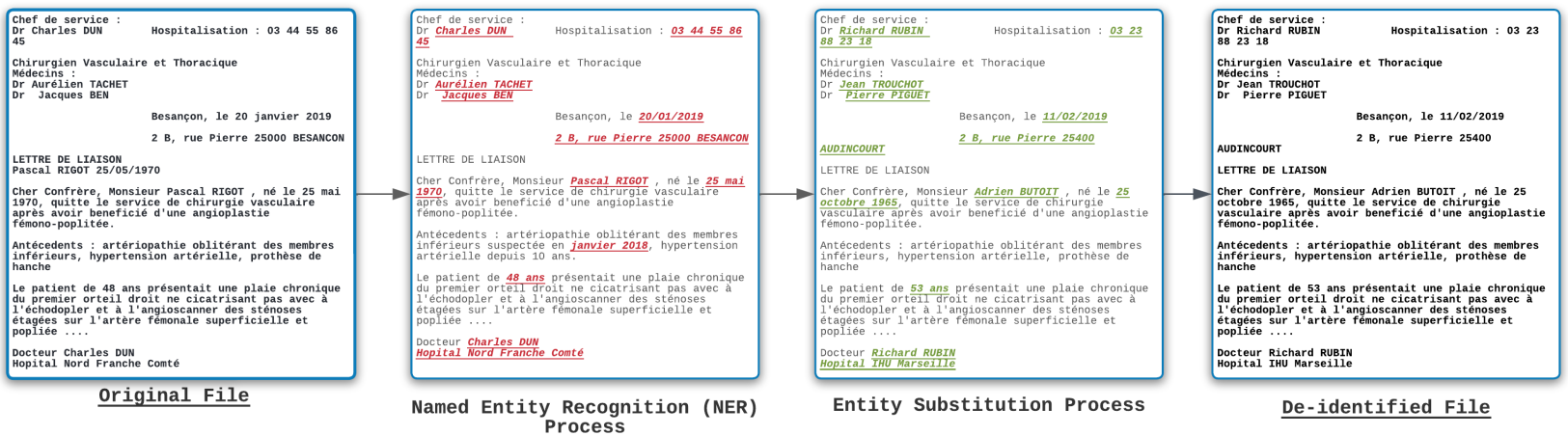
[4] J. C Duchi, M. I Jordan, and M. J Wainwright. Local privacy and statistical minimax rates. In2013 IEEE 54th Annual Symposium onFoundations of Computer Science, pages 429–438. IEEE, 2013.

# Plan

# De-Identification: a Twofold Method

## 2 steps

1. Detection of sensitive information contained in the document
   - Efficiency issue: maximising named entity detection scores
2. Sanitization of detected information
   - Optimisation issue: minimising leakage whilst preserving utility.




## Thread Example:

*Mr. Durand born in Dijon, 40 years old was admitted to the hospital from 12/02/2020 to February 26, 2020 following a road accident in Dijon.*

# Plan

# NER: Searched Entities

## Searched Entities: reduced to HIPAA[5] ones (U.S. Dpt. of Health and Human Services)

| | |
|---|---|
| 1 | Names |
| 2 | All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes |
| 3 | All date elements [...] for dates directly related to an individual including, birth date ... |
| 4,5,6 | Telephone; Fax numbers; E-mail addresses |
| 8 | Medical record numbers |
| 7, 9, 10 | Social security numbers; Health plan beneficiary numbers; Account numbers |
| 11,13 | Certificate/license numbers; Device identifiers and serial numbers |
| 12 | Vehicle identifiers and serial numbers, including license plate numbers |
| 14; 15 | Web universal resource locators (URLs); Internet Protocol (IP) address numbers |
| 16 | Biometric identifiers, including fingerprints and voice prints |
| 17 | Full face photographic images and any comparable images |
| 18 | Any other unique identifying number, feature, or code. |

## Thread Example:

PER LOC AGE

Mr. ***Durand*** born in ***Dijon***, ***40 years old*** was admitted
to the hospital from ***12/02/2020*** to ***February 26, 2020***
following a road accident in ***Dijon*** .

DATE DATE LOC

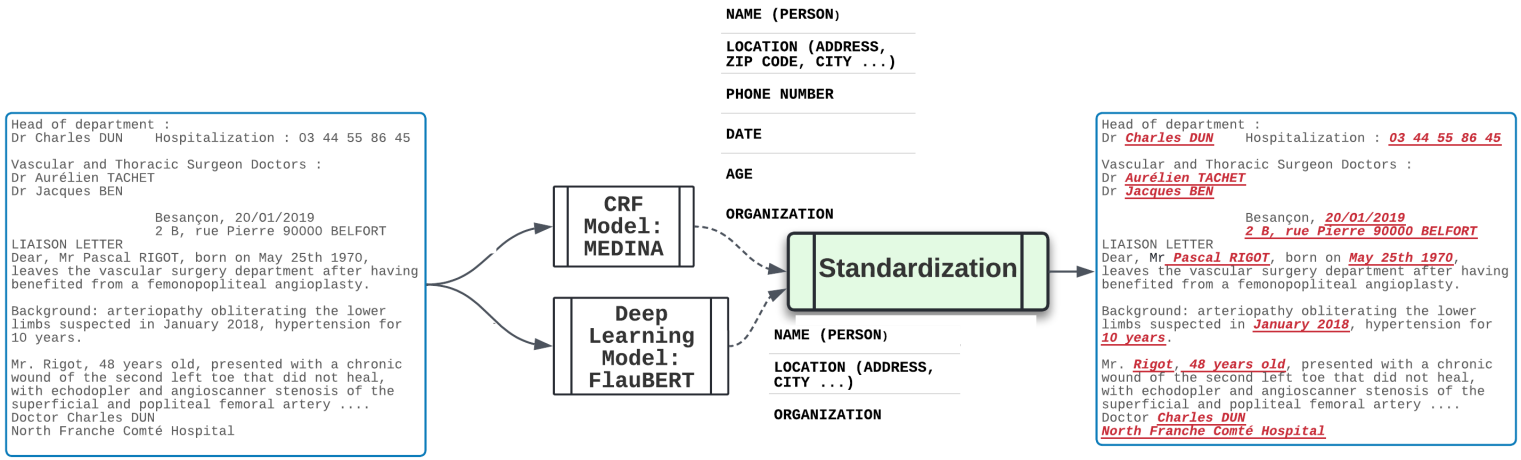
[5] Cohen, I. G., & Mello, M. M. (2018). HIPAA and protecting health information in the 21st century. Jama, 320(3), 231-232.

# NER: issue in French Language

## Issues with French language

Recommandations par PocketEn savoir plus

- WikiNer French[6]: only 4 categories LOC, PER, ORG, and MISC
- MEDINA[78]: rule based and statistical learning, dedicated to French lang.
- **Building of an hybrid system**:




- For ML evaluation: need of a French labelled dataset

[6] Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence, 194, 151-175.

[7] Grouin, C., Griffon, N., & Névéol, A. (2015, September). Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs?. In Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (pp. 31-39).

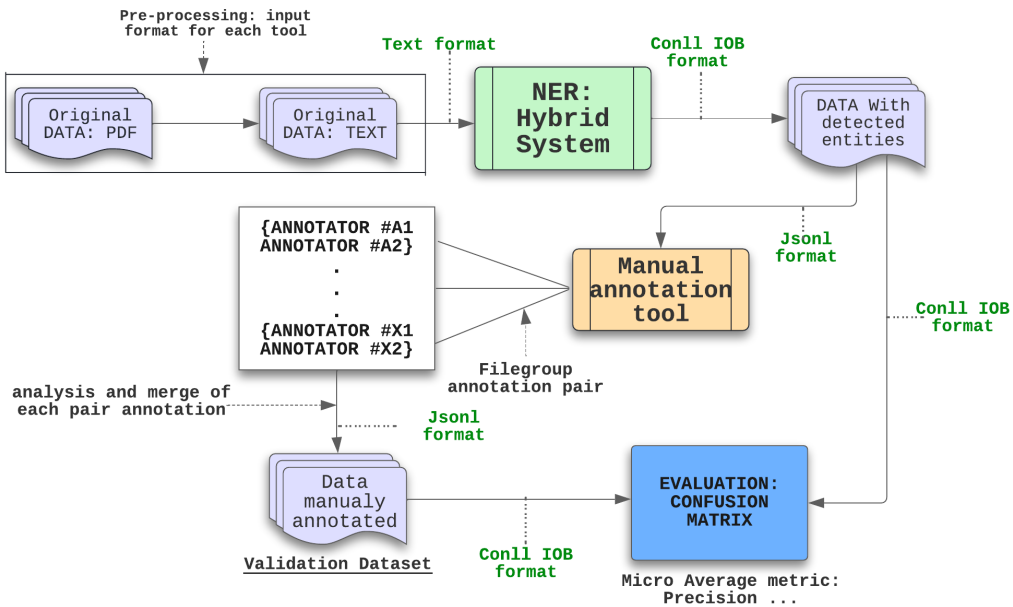
[8] Grouin, C., & Zweigenbaum, P. (2013). Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approaches. In MEDINFO 2013 (pp. 476-480). IOS Press.

# HNFC-NER-EVAL labelled dataset

Methodology: 6 h., @HNFC, 6 people of medical staff

1. Input data: 375 texts of deceased pers., annotated with the hybrid tool
2. Manually annotated by the hospital staff with Doccanno[9]
   - Each annotator completes/corrects errors: "ds. 3 j." vs. "3 x p. j."
   - Merging of pairs of annotation results into a unique annotated file.

→ 9993 sentences, 23829 labels




[9] Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. doccano: Text Annotation Tool for795Human, 2018. Software available from https://github.com/doccano/doccano.

# NER: Evaluating NER Hybrid system[12]

- Training dataset for all ML learning approaches: WikiNer
- Highest recall score: our hybrid proposal
- Micro Average results: hybrid system detects entities most often (highest recall) without neglecting precision (highest F1-score).
- Privacy perspective: issue with miss detection of Dates that are QID...

| Labels | Spacy[10] | | | CamemBERT NER | | | MEDINA | | | FLAUBERT[11] NER | | | Hybrid | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| PERson | 59 | 76.8 | 67 | 89 | 99 | 93.8 | **98.2** | 97.7 | **98.2** | 91.8 | 97.6 | 94.6 | 96.3 | **99.8** | 98 |
| ORGanisat. | 2.2 | 10.9 | 3.6 | 7. | 21.8 | 11.1 | 32.6 | 24.8 | 28.1 | 16.9 | 34.1 | 22.6 | **41.1** | **57.3** | **47.8** |
| LOCation | 40.3 | 11.9 | 18.4 | 46 | 67.2 | 54.6 | **98.8** | 81.1 | 89.1 | 75.7 | 66.3 | 70.7 | 88.4 | **95.8** | **92** |
| Date | | | | | | | 97.7 | 86.6 | 91.9 | | | | **97.7** | **86.7** | **91.9** |
| Age | | | | | | | 91.5 | 66.9 | 77.3 | | | | **91.5** | **66.9** | **77.3** |
| Phone N. | | | | | | | 99.5 | 97.9 | 98.7 | | | | **99.5** | **97.9** | **98.7** |
| Micro Av. | 54.9 | 54.8 | 54.9 | 70.8 | 51.5 | 59.6 | **98.2** | 91.2 | 94.5 | 85.8 | 86.7 | 86.3 | 94.6 | **94.9** | **94.7** |

[10] Montani I spaCy, H. M. (2021). 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[11] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., ... & Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. arXiv preprint arXiv:1912.05372.

[12] Yakini Tchouka and Jean-François Couchot and Maxime Coulmeau and David Laiymani and Philippe Selles and Azzedine Rahmani and Christophe Guyeux 2022. De-Identification of French Unstructured Clinical Notes for Machine Learning Tasks

# Plan

# Entity Substitution: Motivation & Purpose

Depending on the entity relevance to medical task

- Without medical "utility": phone, fax numbers, REF ... $\rightsquigarrow$ pure random approach
- With possible internal links: names $\rightsquigarrow$ random approach whilst preserving affiliation
- With direct impacts on the medical analysis: age, antecedents (DATEs), patient's LOCation

**Thread Example:**

**PER**: Durand $\Rightarrow$ Julien (by random approach)

# Applying $\epsilon$-local differential privacy on DATEs

## Main idea: bounded Laplace Mechanism[13] on intervals

1. Order all normalized dates (d-m-y) $E = [e_0, \ldots, e_n]$, incl. the current one and associate a category (short, medium, long term) to each
2. Compute intervals $I = [e_0 - e_1, \ldots, e_{n-1} - e_n]$ between consecutive dates
3. Bounded Laplace mechanism to each interval $I_i$, w.r.t. category range
4. Reconstitute dates from the current date

## Related Work on dates substitution: uniform shifting of dates

- MIMIC2[14], MIMIC3[15], I2B2[16] datasets, a survey [17] or whitebook[18]

## Attack on HNFC-NER-EVAL dates WITH uniform shifting

- $I = [I_1, \ldots I_{n-2}]$ is NOT modified and is unique in 98% in this dataset

[13] H OLOHAN, Naoise ; A NTONATOS, Spiros ; B RAGHIN, Stefano ; M AC AONGHUSA, Pól : The bounded Laplace mechanism in differential privacy. In arXiv preprint arXiv :1808.10410 (2018)

[14] Douglass, M., Clifford, G. D., Reisner, A., Moody, G. B., & Mark, R. G. (2004, September). Computer-assisted de-identification of free text in the MIMIC2 database. In Computers in Cardiology, 2004 (pp. 341-344). IEEE.

[15] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC3, a freely accessible critical care database. Scientific data, 3(1), 1-9.

[16] `https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/`

[17] Uzuner, Ö., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association, 14(5), 550-563.

# Applying $\epsilon$-local differential privacy on LOCations

Main idea: geo-indistinguishability[19] on coordinates

1. Given a location $Z$ expressed as its polar coordinates
2. Apply a bounded Laplace noise to these coordinates (to reduce sensitivity) and translates this into $Y$ its city name
3. Memoïzation: for each $Z$, use $Y$ in this document to avoid averaging attack

[19] Andrés, M.E.; Bordenabe, N.E.; Chatzikokolakis, K.; Palamidessi, C. Geo-indistinguishability:811Differential privacy for location-based systems. In Proceedings of the Proceedings of the 2013812ACM SIGSAC conference on Computer & communications security, 2013, pp. 901–914

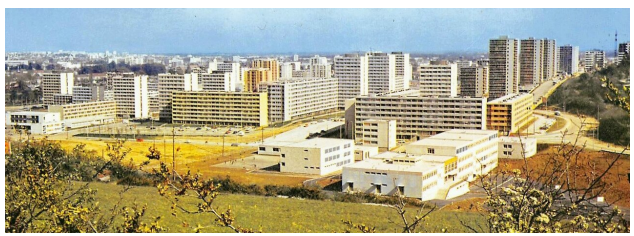
# Analysis of Applying $\epsilon$-LDP

## Motivation for $\epsilon$-local differential privacy

- For an output $o$ and two inputs $v_1$ and $v_2$: both $v_1$ and $v_2$ "may be" the preimage of $o \rightsquigarrow$ strong guaranty for the patient.
- Applying LDP mechanism on Jan. 8, 1942 and March 14, 2018 (birth and death dates of St. Hawking): has to generate approximately the same dates

### Thread Example:

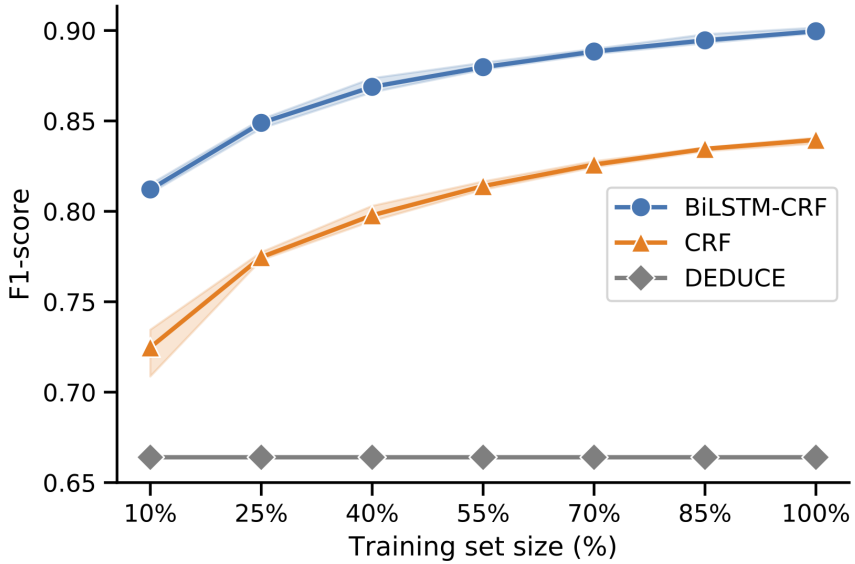
- **DATES**: all are in the long term category (with large sensitivity)
  - February 26, 2020 $\Rightarrow$ Oct. 05 2020
  - 12/02/2020 $\Rightarrow$ 23/06/2015 (very long stay: utility?)
  - 40 years old $\Rightarrow$ 30 years old
- **LOC**: a regional capital DIJON $\Rightarrow$ a charming village BEZE (with completely opposite epidemiological data)



# Plan

# Deep Learning vs other models in NLP

## Comparing NER scores of Dutch medical records de-identification[20]




- Combining BiLSTM-CRF in deidentifications is accurate, but it always remains errors

## Metrics on GLUE[21] benchmark when BERT[2] was introduced

| System | MNLI-(m/mm) 392k | QQP 363k | QNLI 108k | SST-2 67k | CoLA 8.5k | STS-B 5.7k | MRPC 3.5k | RTE 2.5k | **Average** - |
|---|---|---|---|---|---|---|---|---|---|
| Pre-OpenAI SOTA | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 |
| BiLSTM+ELMo+Attn | 76.4/76.1 | 64.8 | 79.8 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 |
| OpenAI GPT | 82.1/81.4 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 |
| $\text{BERT}_{\text{BASE}}$ | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| $\text{BERT}_{\text{LARGE}}$ | **86.7/85.9** | **72.1** | **92.7** | **94.9** | **60.5** | **86.5** | **89.3** | **70.1** | **82.1** |

- Outperforms all other approaches
- Requires larger training dataset

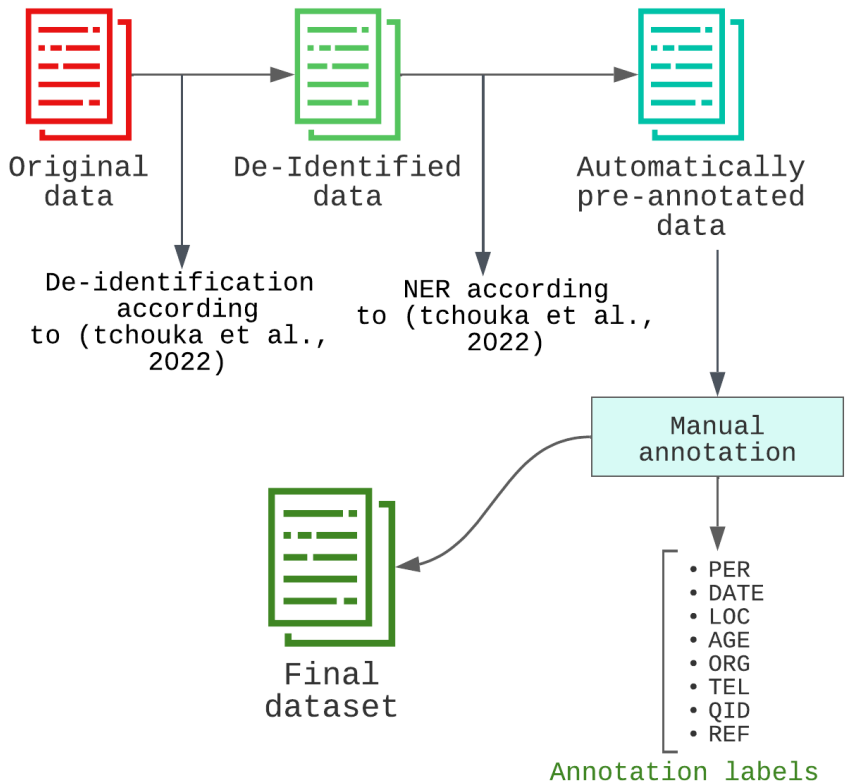
[20] Trienes, J., Trieschnigg, D., Seifert, C., & Hiemstra, D. (2020). Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records. arXiv preprint arXiv:2001.05714.

[21] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.

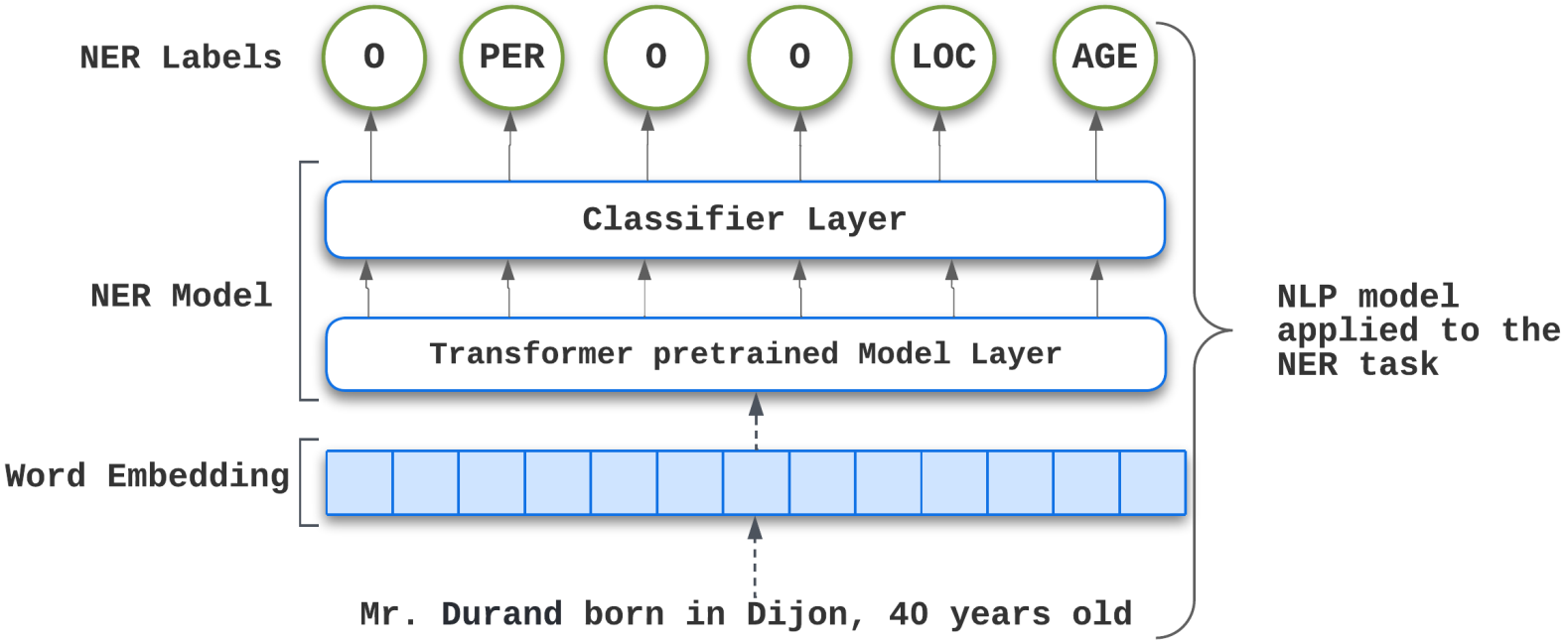
# HNFC-NER-TRAIN labelled dataset for DL training

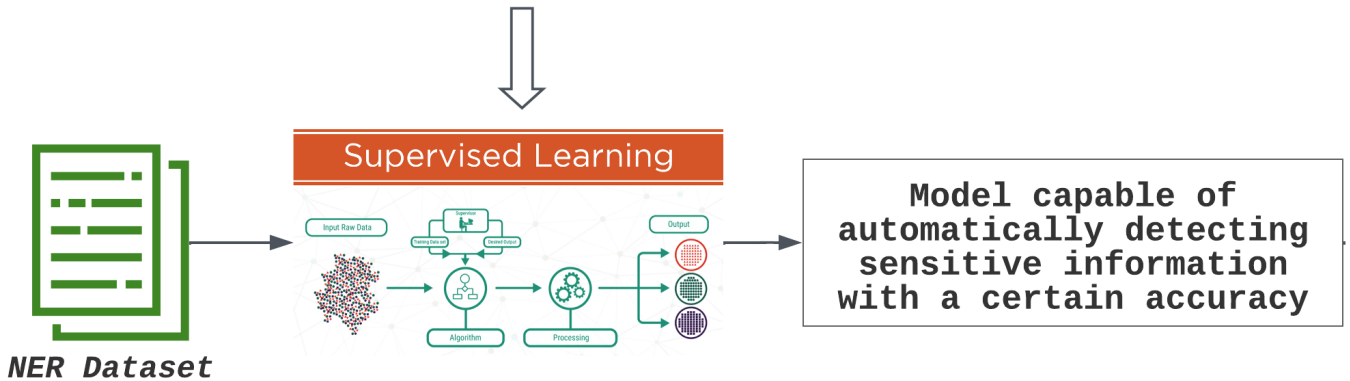
Methodology: 25 h., @HNFC, 1 people.

1. Input data: 1500 texts (14925 sentences) of deceased pers., first deidentified and next pre-annotated by previous hybrid approach
2. Manually annotated @HNFC with Doccanno again



# FLUBERT NER Model Architecture



# NER results

## Improved results for almost all metrics

| Methods | Hybrid Syst[12] | | | PROPOSAL | | | Denoncourt System (RNN)[22] | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | HNFC-NER-EVAL | | | | | | i2b2 | | |
| Metrics | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| PER | 96.3 | **99.8** | 98 | 97.2 | 98.9 | 98 | **98.2** | 99.1 | **98.6** |
| ORG | 41.1 | 57.3 | 47.8 | 90 | 51 | 65.6 | **92.9** | **71.4** | **80.7** |
| LOC | 88.4 | **95.8** | 92 | **99.4** | 94.4 | **96.9** | 95.9 | 95.7 | 95.8 |
| DATE | 97.7 | 86.7 | 91.9 | **99.2** | 95.7 | 97.4 | 99 | **99.5** | **99.2** |
| AGE | 91.5 | 66.9 | 77.3 | 98.2 | 91.8 | 95 | **98.9** | **97.6** | **98.2** |
| TEL | **99.5** | 97.9 | 98.7 | 99.4 | **99.8** | **99.6** | 98.7 | 99.7 | 99.2 |
| REF | | - | | 96.1 | 79.5 | 87 | | - | |
| Micro av. | 94.6 | 94.9 | 94.7 | **98.5** | 96.4 | 97.4 | 98.3 | **98.5** | **98.4** |

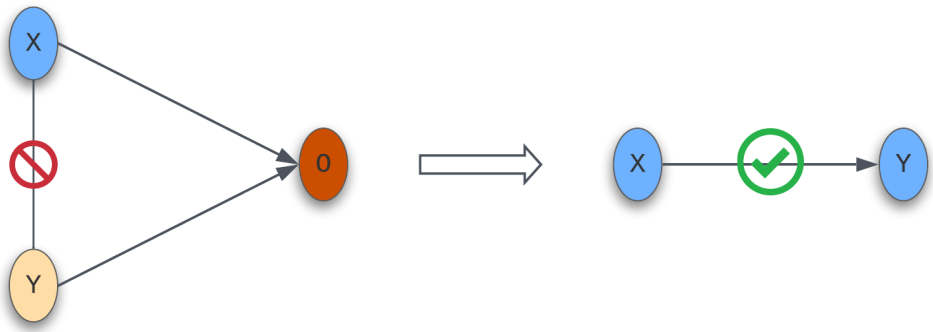
- But still not as strong as English-language results

[22] F. Dernoncourt and J. Lee and O Uzuner and P. Szolovits 2016. De-identification of Patient Notes with Recurrent Neural Networks

# Plan

# $\epsilon.d$-privacy[23]: a $\epsilon$-LDP relaxation

## Motivation




- 2 distant elements: should not be confused by a random mechanism
- Need of mechanism taking into account the distance between elements

## Definition ($\epsilon.d$-privacy for a mechanism $\mathcal{M}$)

$\mathcal{M}$ verifies the $\epsilon.d$-privacy if, $\forall v_1, v_2 \in \text{Domain}(\mathcal{M}), \forall o \in \text{Image}(\mathcal{M})$,
$\Pr[\mathcal{M}(v_1) = o] \leq e^{\epsilon.d(v_1,v_2)} \cdot \Pr[\mathcal{M}(v_2) = o]$, with $d$ metric on $\text{Domain}(\mathcal{M})$

[23] Chatzikokolakis, Konstantinos, et al. "Broadening the scope of differential privacy using metrics." International Symposium on Privacy Enhancing Technologies Symposium. Springer, Berlin, Heidelberg, 2013.

# Applying $\epsilon.d$-privacy on DATEs

## Date Value

- Date conversion by unit (year, day, month, ...):
  $E = [e_0, e_1, \ldots, e_n] \Rightarrow I = [I_1, I_2, \ldots, I_n]$
- Dates becoming numeric values $\rightsquigarrow d(d_1, d_2) = |d_1 - d_2|$

## Application to

- $\epsilon.d$-privacy on distance $|x| \Rightarrow$ Laplace distribution centered at 0 with scale parameter $1/\epsilon$
- With some post-processing (rounding interval duration)

**Thread Example:**

- February 26, 2020 $\Rightarrow$ March 01, 2020
- 12/02/2020 $\Rightarrow$ 20/02/2020
- 40 years old $\Rightarrow$ 37 years old

# Applying $\epsilon$-d privacy on LOCations

## Distance between locations

| city | overall population | cancer incidence rate | stroke | distance | scores | normalized distribution |
|---|---|---|---|---|---|---|
| DIJON | 160204 | 182.252004 | 273.184785 | 0.000000 | 1.000000 | 0.133468 |
| BESANCON | 119249 | 134.135495 | 218.375283 | 0.418721 | 0.581279 | 0.120203 |
| CHALON SUR SAONE | 46603 | 52.730489 | 108.706972 | 1.170695 | -0.170695 | 0.099602 |
| DOLE | 24606 | 57.437117 | 55.290112 | 1.349742 | -0.349742 | 0.095242 |
| LONS LE SAUNIER | 18023 | 42.070599 | 40.497996 | 1.450857 | -0.450857 | 0.092865 |
| LE CREUSOT | 21935 | 24.819073 | 51.165964 | 1.466909 | -0.466909 | 0.092493 |
| VESOUL | 15728 | 42.069461 | 33.302482 | 1.475195 | -0.475195 | 0.092301 |
| BEAUNE | 21747 | 24.739921 | 37.083653 | 1.497015 | -0.497015 | 0.091799 |
| MONTCEAU LES MINES | 18789 | 21.259429 | 43.827550 | 1.504867 | -0.504867 | 0.091619 |

- Epidemiological data of each location: as a vector, further normalized

## Randomization: like exponential mechanism

- Scoring function $U(j, i) = 1 - d(i, j)$
- Substitutes limited to the $k$ closest locations w.r.t. the distribution $P_j = [a.e^{\epsilon U(j,i_1)}, \ldots, a.e^{\epsilon U(j,i_k)}, 0, \ldots, 0]$

**Thread Example:**

- **LOC**: Dijon $\Rightarrow$ Besançon

# Result on the Thread Example

## Thread Example:

Mr. Durand born in Dijon, 40 years old was admitted to the hospital from 12/02/2020 to February 26, 2020 following a road accident in Dijon.

↓

Mr. Julien born in Besançon, 37 years old was admitted to the hospital from 20/02/2020 to March 01, 2020 following a road accident in Besançon.

**De-Identification Tool**

# Plan

# A short history to be continued

- C. Grouin's work on de-identification in French[24]
- i2b2[16], MIMIC3[15] public dataset de-identified by Stubbs[25]
- Recent work on de-identification[26] with GPT-4

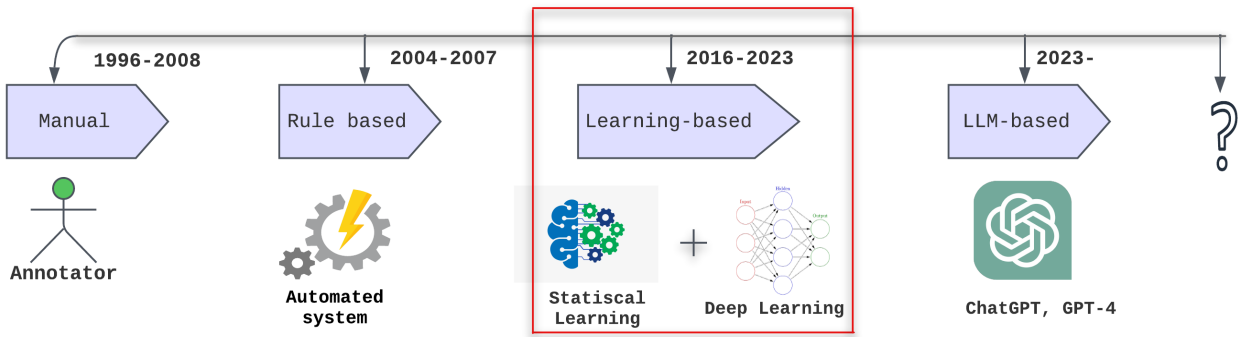



Figure: An history of de-identification[26]

[24] Grouin, C. and Névéol, A. 2014. De-identification of clinical notes in French: towards a protocol for reference corpus development

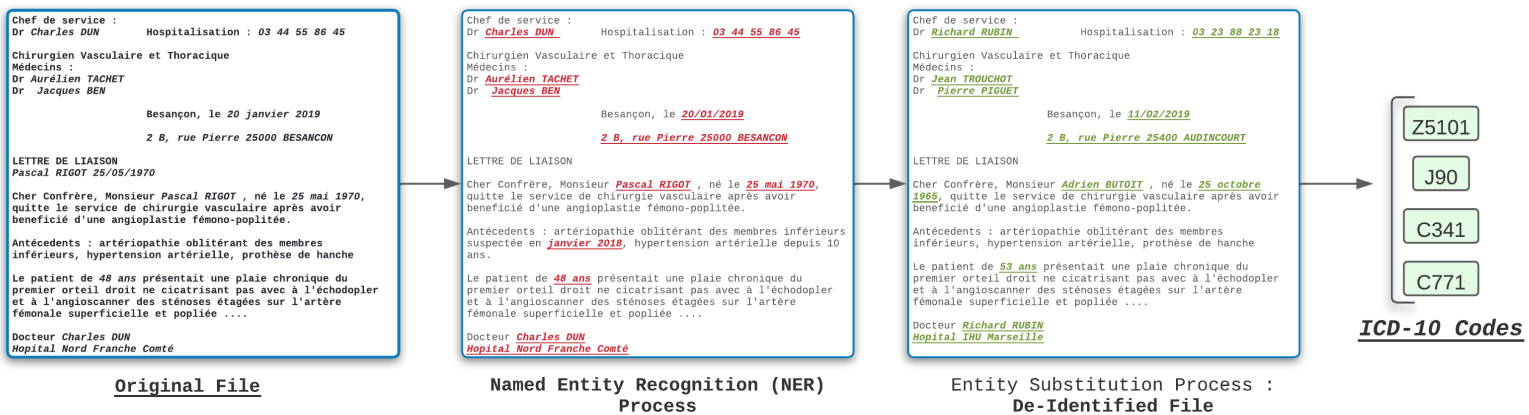
[25] Stubbs, Amber and Uzuner, Özlem and Kotfila, Christopher and Goldstein, Ira and Szolovits, Peter 2015. Challenges in synthesizing surrogate PHI in narrative EMRs

[26] Liu, Zhengliang and Yu, Xiaowei and Zhang, Lu and Wu, Zihao and Cao, Chao and Dai, Haixing and Zhao, Lin and Liu, Wei and Shen, Dinggang and Li, Quanzheng and others 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4

# Plan

# ICD-10 Codes

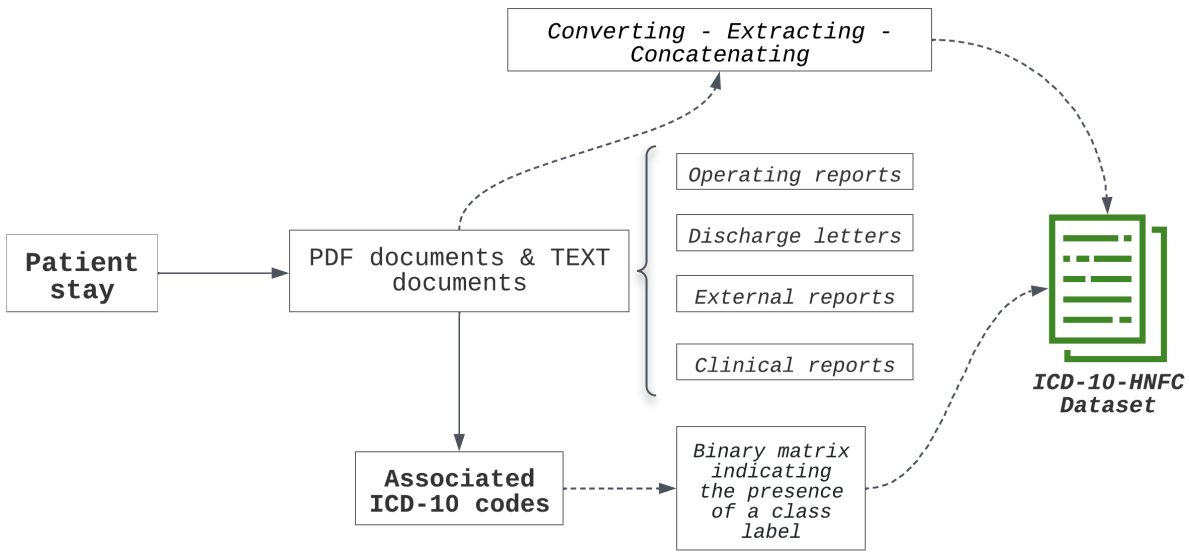
- ICD-10 (International Classification of Diseases, Tenth Revision) codes : standardized system used for classifying and coding diseases, injuries, and other health-related conditions
- Assigned to medical diagnoses and procedures to facilitate accurate and consistent recording and reporting of health information
- Each healthcare stay: manual summarized into ICD-10 codes (stat., remuneration)
- Computing: a multi-label classification of unstructured data



# ICD-10-HNFC dataset for multi-label classification

## Very private dataset, @HNFC

- Input data: 56,014 patient stays composed of medical texts with their associated ICD-10 codes
- Output: 56,014 (very long) lines, concatenation results and their associated binary vector of labels
- Second output: same text and ICD-10 codes group by families (class reduction)



# ICD-10-HNFC dataset : challenging metrics

## Descriptive statistics of ICD-10-HNFC dataset

| | Dataset | Dataset with class reduction |
|---|---|---|
| Documents | 56014 | - |
| Tokens | 41868993 | - |
| Average sequence length | 747 | - |
| Total ICD codes | 416125 | 415830 |
| Unique ICD codes | 6160 | 1564 |
| Codes with less than 10 examples | 3722 | 523 |
| Codes with 100 examples or more | 641 | 471 |

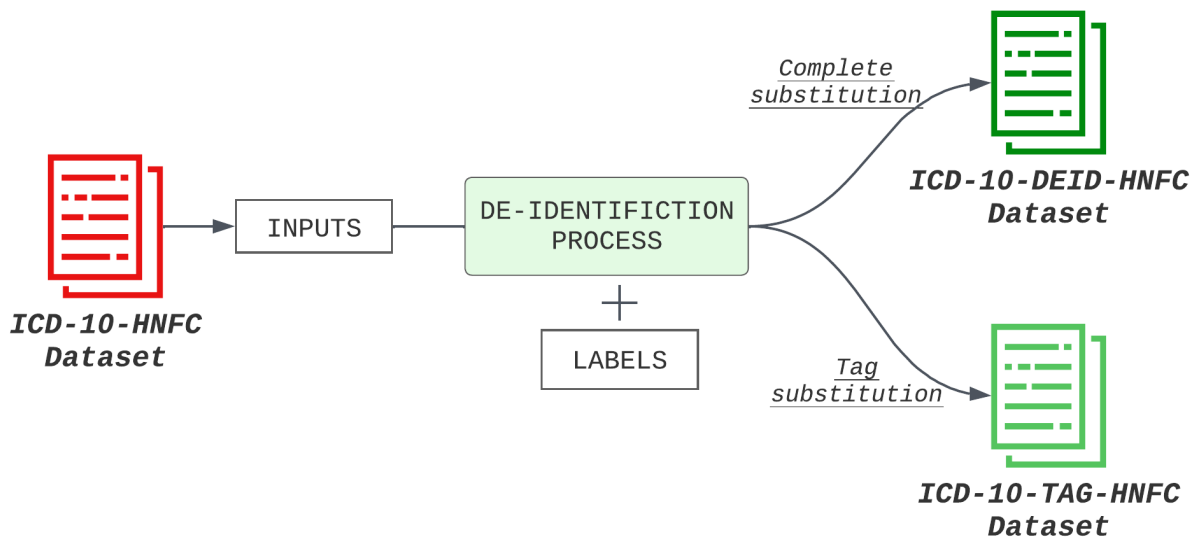
## Two issues in ICD-10 codes association

1. Input patient file : usually a long sequence:
   - Average sequence length (747) $>$ maximum input size for Transformers (512): scalability issue
2. Large number of different codes, labels, but sparse
   - 6160 unique ICD codes, 3722 of whom have only been less than 10 times: scalability and sparsability issue

# ICD-10-DEID-HNFC (ICD-10-TAG-HNFC): working dataset

Two de-identified datasets, @HNFC, we can work with

- Input data: ICD-10-HNFC dataset
- Output 1: ICD-10-DEID-HNFC with aforementioned de-identification approach
- Output 2: ICD-10-TAG-HNFC with tag only substitution (baseline)
- 10,000 lines are removed (everywhere) due to errors in date format or not found locations in optimal de-identification

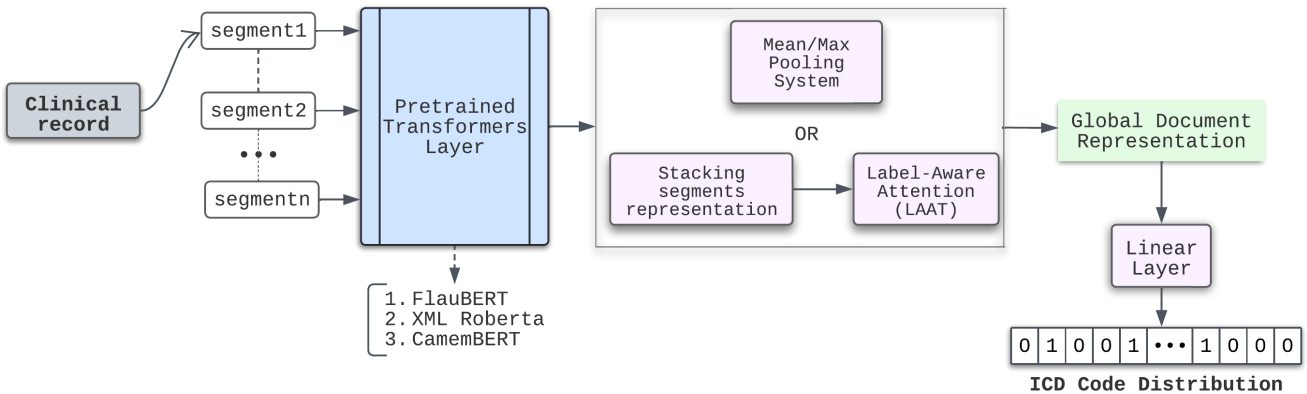


# ICD-10 codes association model

## Approach with FLAUBERT

- Long sequence processing: Hierarchical Transformers[27]
  1. Document divided into segments → representation of each segment with pre-trained Transformers layer
  2. Aggregation ⇝ Document representation
- Large and sparse label set: Label-Aware Attention mechanism (LAAT)[28]
  - Labels are integrated into the document representation

## Model Architecture

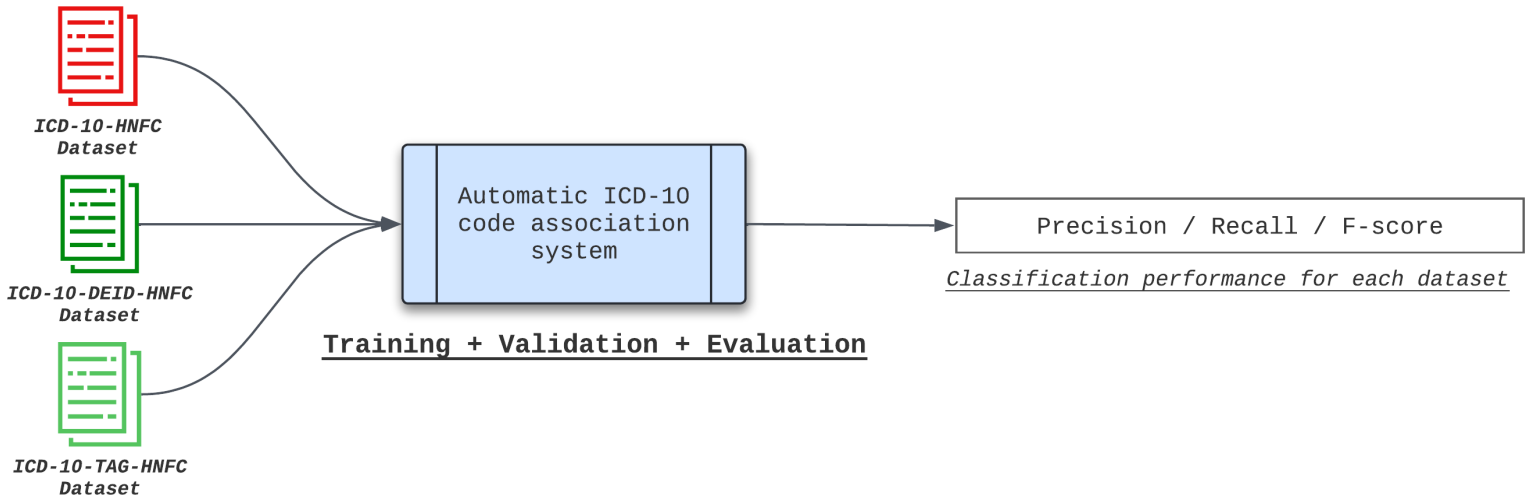



[27] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019, December). Hierarchical transformers for long document classification. In 2019 IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 838-844). IEEE.

[28] Huang, C. W., Tsai, S. C., & Chen, Y. N. (2022). PLM-ICD: automatic ICD coding with pretrained language models. arXiv preprint arXiv:2207.05289.

# Evaluating ICD-10 codes association on (de-identified) datasets

## Automatic association of ICD-10 codes on different corpora (de-identified or not)




## Results on the evaluation dataset

| Dataset | Labels | Precision | Recall | $F_1$-score |
|---|---|---|---|---|
| ICD-10-TAG-HNFC | 6160 | 0.43 | 0.41 | 0.42 |
| ICD-10-DEID-HNFC | | 0.44 | 0.43 | 0.44 |
| ICD-10-HNFC | | 0.47 | 0.46 | 0.47 |

- ICD-10-DEID-HNFC: enabled us to prototype the whole ML approach
- ICD-10-DEID-HNFC vs ICD-10-TAG-HNFC: most accurate, close to original ones

# State of the art of ICD-10 codes association

## Experimental results

| Models | Language | Dataset | Labels | $F_1$-score |
|---|---|---|---|---|
| PLM-ICD[29] | English | MIMIC2<br>MIMIC3 | 5,031<br>8,922 | 0.5<br>**0.59** |
| Bouzille[30] | French | own dataset | 6,116<br>1,549 | 0.39<br>0.52 |
| | | ICD-10-HNFC | 6,161<br>1,564 | 0.27<br>0.35 |
| **PROPOSAL** | | | 6,161<br>1,564 | **0.45**<br>**0.55** |

- Bouzille: same parameter than the ones in[30]
- All codes (Bouzille and ours) will be (very soon) on github
- State-of-the-art ICD-10 codes association model[31] in French language

[29] Huang, C. W., Tsai, S. C., & Chen, Y. N. (2022). PLM-ICD: automatic ICD coding with pretrained language models. arXiv preprint arXiv:2207.05289.

[30] BOUZILLE, G., & GRABAR, N. (2020). Supervised learning for the ICD-10 coding of French clinical narratives. Digital Personalized Health and Medicine: Proceedings of MIE 2020, 270, 427.

[31] Tchouka, Y., Couchot, J. F., Laiymani, D., Selles, P., & Rahmani, A. (2023). Automatic ICD-10 Code Association: A Challenging Task on French Clinical Texts. arXiv preprint arXiv:2304.02886.

# Plan

# Conclusion

## Contributions on De-identification

- Complete accurate differentially private de-identification method
  - State-of-the-art NER model for de-identification in French language
- Substitution method that combines **utility** and **safety**
  - Not location-specific Method
  - Solution available on github[32]

## Contributions on ICD-10 codes association task

- Deep learning system that combines the latest advances in Natural Language Processing
- State-of-the-art ICD-10 codes association model in French language

## Future work

- Using this deidentification method to provide a clinicalBERT à la française
- Evaluating the security of the approach against membership inference attack

[32] Surrogate Generation in De-identification. 2022