Prototyping on sensitive medical data: possible thanks to de-identification verifying differential privacy.

Jean-François Couchot[1]

[1]Université de Franche-Comté, FEMTO-ST, France

Séminaire RDI BMB /LMB

Plan

Outline

Legal Context of De-Identifying Clinical Textual Documents

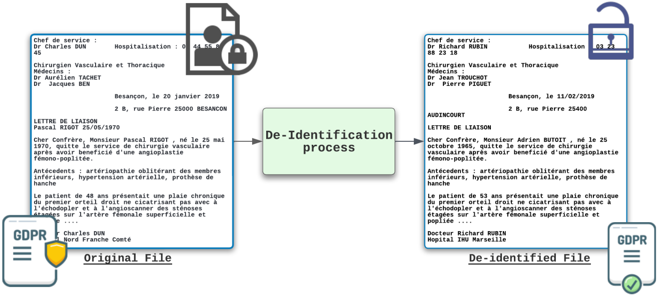
Considered Data Type

- Unstructured data: Clinical textual documents containing information such as names, ages, and locations.
  - Natural Language Processing (NLP) task.
- Excludes images or tabular data.

Legal Requirements

- Enable medical data accessibility for researchers while safeguarding patient privacy.
- Legal requirements mandated by legislation before data sharing:
  - GDPR: Delete any data that could identify an individual, which necessitates de-identification.
  - HIPAA: Provides a list of 18 attributes to be removed from medical documents, making de-identification more explicit.

De-Identification: Global Overview



Researchers with De-Identified Data Can

- Provide models for other medical tasks (e.g., clinicalBERT[1], a BERT[2] specialization).
- Apply further NLP tasks, such as text summarization or, in this case, multi-label classification tasks (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-Identification with Differential Privacy

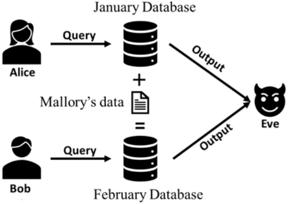
What is differential privacy? See next slides.

Plan

Plan

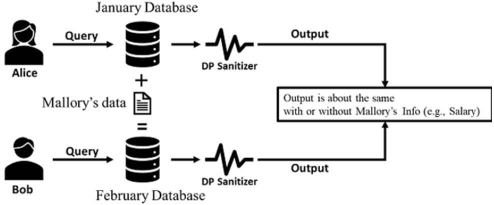
Example of Queries on Neighboring Databases[3]

Without Differential Privacy



- Monthly query: (#employees, average salary).
- Result: $\{\text{Jan} : (100, \$55,000), \text{Feb} : (101, \$56,000)\}$.
- Suppl. knowledge: 0 output + Mallory in February.
- $\rightsquigarrow$ Mallory's salary: $156,000.

With Differential Privacy



- Same queries, same additional knowledge.
- Sanitized results: $\{\text{Jan} : (102, \$55,551),$ $\text{Feb} : (97, \$55,975)\}$.
- Mallory's salary?

[3] Privacy-Preserving Machine Learning. Manning Early Access Program Publications, 2021.

Key Ideas

Intuition for Two Neighboring Databases D_1 and D_2

- Results (aggregated, statistical, etc.) are close.
- $\Leftrightarrow$ "Probabilities" on $\mathcal{M}(D_1)$ and $\mathcal{M}(D_2)$ are nearly equal (up to ϵ).

Why Differential Privacy?

- Private data: desire to have little impact on results.
- $\rightsquigarrow$ Difficult to distinguish if a particular individual "participates or not."
- $\rightsquigarrow$ Data owner is less concerned about sharing their data.

Plan

Formalization of Differential Privacy[4]

Definition (ϵ-Differential Privacy (DP))

ϵ-Differential Privacy (DP): So let $\epsilon \in \mathbb{R}^+$. The non-deterministic probabilistic algorithm $\mathcal{M}$ satisfies ϵ-Differential Privacy if

$$
\begin{array}{ll}
\forall D_1, D_2 \in \mathbb{N}^{|\mathcal{X}|} \text{ such that } \|D_1 - D_2\|_1 = 1, & (D_1,\ D_2\text{: neighboring databases}) \\
\forall R \text{ such that } R \subseteq \mathcal{M}(\mathbb{N}^{|\mathcal{X}|}), & (\text{for any output of the algorithm}) \\
\Pr[\mathcal{M}(D_1) \in R] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in R] & (\text{if } \epsilon \text{ is small, } e^{\epsilon} \approx 1 + \epsilon)
\end{array}
$$

Budget of Leakage $\epsilon \in \mathbb{R}^+$: Allowed Deviation, Permitted Leakage

- $\Pr[\mathcal{M}(D_1) \in R] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in R]$: results are approximately equal (but not necessarily) with or without the data of one person.
- $\epsilon = 0$: No deviation is allowed (all outputs are equal with or without the data of one person), data is perfectly protected (but less useful).
- Small vs. large ϵ: It depends on the amount of permitted leakage.

[4] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265-284). Springer, Berlin, Heidelberg.

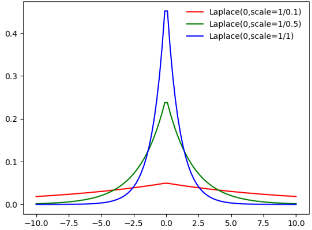
Plan

Query Q_1: Number of Employees in the Database

Objectives, Data, Idea

- Publish the number of employees with an ϵ-DP mechanism.
- $Q_1(D_{\text{Jan}}) = 100$, $Q_1(D_{\text{Feb}}) = 101$, etc.
- Add Laplace noise centered at 0 depending on ϵ.

Implementation: Laplace Noise Centered at 0, $\mathcal{M}_L(D) = Q_1(D) + v$, $v \sim Lap(0, \epsilon^{-1})$

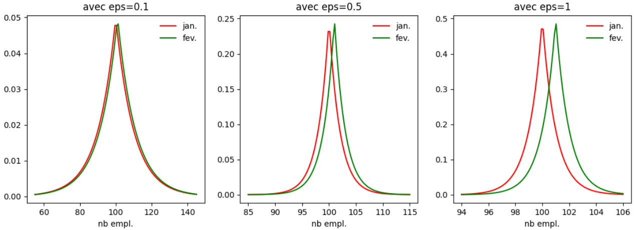


Query Q_1: Number of Employees in the Database

Objectives, Data, Idea

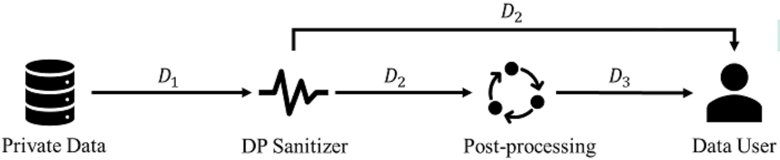
- Publish the number of employees with an ϵ-DP mechanism.
- $Q_1(D_{\mathrm{Jan}}) = 100$, $Q_1(D_{\mathrm{Feb}}) = 101$, etc.
- Add Laplace noise centered at 0 depending on ϵ.

Implementation: Laplace Noise Centered at 0, $\mathcal{M}_L(D) = Q_1(D) + v$, $v \sim Lap(0, \epsilon^{-1})$



Robustness to Post-Processing

Intuition for a Database D_1[5]



Interpretations

- Post-processing if seen as a subsequent algorithm (e.g., removing outliers): only the DP algorithm needs to be considered carefully.
- Post-processing seen as an attack by an adversary: they can incorporate as much auxiliary information as they want; the privacy guarantee remains valid.

Theorem (Post-Processing of an ϵ-DP Mechanism)

For any function $f : \mathcal{M}(\mathbb{N}^{|\mathcal{X}|}) \rightarrow f(\mathcal{M}(\mathbb{N}^{|\mathcal{X}|})$, $f(\mathcal{M})$ *is also* ϵ*-DP.*

Direct application

- Any sanitized real data: can subsequently be rounded to the nearest integer.

Composition of Sequential Leaks

Sequences of Leaks

- It is common to query the same database iteratively (e.g., employee count in January, February, etc.).
- Each query corresponds to a data leak, and we want to find the total leakage for a sequence of leaks with ϵ_1 and ϵ_2.

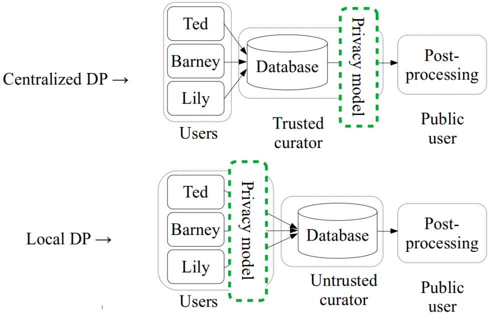
Theorem (Sequential Composition of ϵ-DP Mechanisms)

If $\mathcal{M}_1$ and $\mathcal{M}_2$ operate on non-disjoint sets, $\mathcal{M}_{1,2}$ is $\epsilon_1 + \epsilon_2$-DP.

Outline

Motivations

In Visual Terms



Differential Privacy (DP) vs. Local Differential Privacy (LDP)

- Trust required in the Database Management System (DBMS).
- Optimal noise per query.

- Individual noise for all post-processing (e.g., Machine Learning).
- Unnecessary trust in the DBMS.

Definition[6] and Properties

Definition of ϵ-Local Differential Privacy (ϵ-LDP)

- $\mathcal{X}$: the set of possible input values.
- $\epsilon \in \mathbb{R}^+$: privacy budget.
- $\mathcal{M}$: non-deterministic probabilistic algorithm respects ϵ-Local Differential Privacy if

$$\begin{array}{ll} \forall x_1, x_2 \in \mathcal{X} & (x_1 \text{ and } x_2 \text{ are two input data points}) \\ \forall y \text{ s.t. } y \in \mathcal{M}(\mathcal{X}), & (\text{for any output } y \text{ of the algorithm}) \\ \Pr[\mathcal{M}(x_1) = y] \leq e^{\epsilon} \Pr[\mathcal{M}(x_2) = y] & \end{array}$$

Properties Similar to DP

- Robustness to post-processing.
- Combining two mechanisms ϵ_1-LDP and ϵ_2-LDP results in $\epsilon_1 + \epsilon_2$-LDP.

[6] Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013, October). Local privacy and statistical minimax rates. In 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (pp. 429-438). IEEE.

Motivation: Dealing with Sensitive Data[8]

Table with a Single Binary Attribute: Q_1 = "Have you ever cheated?"

- Embarrassment: temptation for a student not to respond honestly.

Randomization according to Warner[7]

- Each student flips two coins $\{\text{Heads}, \text{Tails}\}$ without revealing the two successive results t_1 and t_2.
- Addition of question Q_2: "Is t_2 equal to Heads?"
  - If t_1 is Heads, the student responds honestly to question Q_1.
  - Otherwise ($t_1 = \text{Tails}$), the student responds honestly to question Q_2.

Analysis of the Extension

- Partially random response: We do not know if an individual's "yes" response originates from dishonesty or a Heads result on the second flip.
- Enhanced honesty of the student: It is the student who modifies their data.

[7] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309), 63-69.

[8] https://fr.coursera.org/lecture/stanford-statistics/warners-randomized-response-model-ck65q

LDP on Continuous Data: Laplace Mechanism Again

Continuous Interval of Width Δ: Bounded Laplace Mechanism $\mathcal{M}_{Lb}$

- $\mathcal{M}_{Lb}(x) = x + v$ s.t. $v \sim Lap(\frac{\Delta}{\epsilon})$
- If $x + v$ falls outside the interval, apply $\mathcal{M}_{Lb}$ again.

Outline

$\epsilon.d$-**Privacy**[9]

Motivation

- (L)DP: it's challenging to determine the origin of a given output.
- 2 data points, far apart $\rightsquigarrow$ may produce the same output.
- Relevance when dealing with a large data space (e.g., centuries, the entire Earth)?
- Introduction of the concept of distance between data points in the probability constraint.

Definition of $\epsilon.d$-Privacy

- $\mathcal{X}$: the set of possible input values, equipped with a metric d.
- $\mathcal{M}$: non-deterministic probabilistic algorithm that adheres to $\epsilon.d$-privacy if

$$\forall x_1, x_2 \in \mathcal{X} \qquad \text{(x_1 and x_2 are two input data points)}$$
$$\forall y \text{ s.t. } y \in \mathcal{M}(\mathcal{X}), \qquad \text{(for any output y of the algorithm)}$$
$$\Pr[\mathcal{M}(x_1) = y] \leq e^{\epsilon.d(x_1,x_2)} \Pr[\mathcal{M}(x_2) = y]$$

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

Plan

Introduction to De-Identification

Introduction to Differential Privacy

De-Identification: an Incremental Approach with Differential Privacy
- De-Identification: A Twofold Method
- Named Entity Recognition, First Attempt
- Entity Substitution: First Attempt
- Named Entity Recognition, Continuation
- Entity Substitution, Continuation

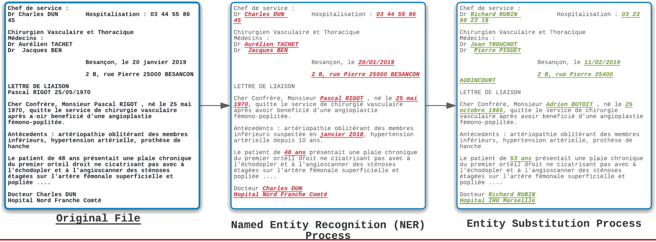
Application of de-identification to ICD-10 codes association

Conclusion

De-Identification: A Twofold Method

Two Steps

1. Detection of sensitive information contained in the document.
   - Efficiency issue: Maximizing named entity detection scores.
2. Sanitization of detected information.
   - Optimization issue: Minimizing leakage while 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 Categories (U.S. Department 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 LOC DATE

NER: Issue in French Language

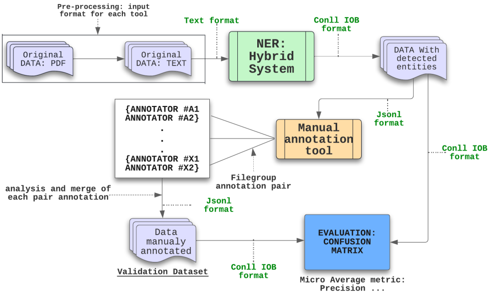
Issues with the French Language

- Limited entity categories in French NER datasets, e.g., only four categories in WikiNer.
- Rule-based and statistical learning approaches in MEDINA and rule-based systems.
- Development of a hybrid system to address these limitations.
- Need for a labeled French dataset for machine learning evaluation.

HNFC-NER-EVAL Labeled Dataset

Methodology: 6 hours, 6 people of the medical staff, @HNFC

1. Input data: 375 texts of deceased persons, annotated with the hybrid tool.
2. Manually annotated by the hospital staff using Doccanno.
   - Each annotator completes/corrects errors, e.g., "ds. 3 j." vs. "3 x p. j."
   - Merging of pairs of annotation results into a unique annotated file.
3. Result: 9,993 sentences, 23,829 labels.



Outline

Entity Substitution: Motivation and Purpose

Dependent on the Entity's Relevance to Medical Tasks

- Entities with no medical utility, such as phone numbers, fax numbers, and references: A pure random approach is applied.
- Entities with possible internal links, like names: A random approach is applied while preserving the affiliation.
- Entities with direct impacts on medical analysis, such as age, antecedents (dates), and the patient's location.

Thread Example:

PER: Durand $\Rightarrow$ Julien (via a random approach)

Applying ϵ-Local Differential Privacy to Dates

Main Idea: Bounded Laplace Mechanism on Intervals[10]

1. Order all normalized dates (day-month-year) $E = [e_0, \ldots, e_n]$, including the current date, 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. Apply the bounded Laplace mechanism to each interval I_i, considering the category range.
4. Reconstruct dates from the current date.

Related Work on Date Substitution: Uniform Shifting of Dates

- MIMIC2[11], MIMIC3[12], I2B2[13] datasets.

Attack on HNFC-NER-EVAL Dates with Uniform Shifting

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

---

[10] Holohan, Naoise; Antonatos, Spiros; Braghin, Stefano; Mac Aonghusa, Pól: The Bounded Laplace Mechanism in Differential Privacy. In arXiv preprint arXiv:1808.10410 (2018)

[11] 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.

[12] 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.

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

Applying ϵ-Local Differential Privacy to Locations

Main Idea: Geo-Indistinguishability on Coordinates[14]

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

[14] Andrés, M.E.; Bordenabe, N.E.; Chatzikokolakis, K.; Palamidessi, C. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In Proceedings of the Proceedings of the 2013 ACM SIGSAC conference on Computer & Communications Security, 2013, pp. 901–914

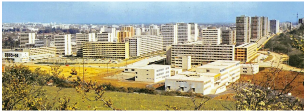
Analysis of Applying ϵ-Local Differential Privacy

Motivation for ϵ-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, providing a strong guarantee for the patient's privacy.
- 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:

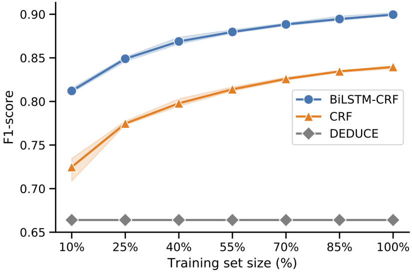
- **DATES**: All are in the long-term category (with large sensitivity).
  - February 26, 2020 $\Rightarrow$ Oct. 05, 2020
  - 12/02/2020 $\Rightarrow$ 23/06/2015 (very long stay: utility?)
  - 40 years old $\Rightarrow$ 30 years old
- **LOC**: A regional capital DIJON $\Rightarrow$ a charming village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



Outline

Deep Learning vs. Other Models in NLP

Comparing NER Scores for Dutch Medical Records De-Identification[15]



- Combining BiLSTM-CRF for de-identification is accurate, but errors still occur.

Metrics on GLUE[16] benchmark when BERT[2] was introduced

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
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 a larger training dataset.

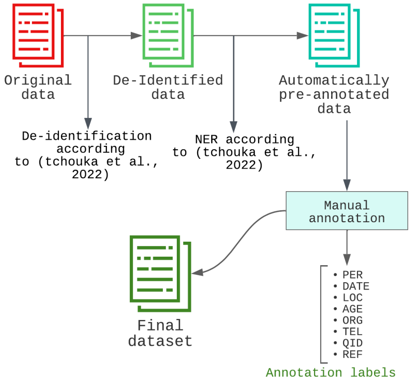
[15] Trienes, J., Trieschnigg, D., Seifert, C., & Hiemstra, D. (2020). Comparing Rule-based, Feature-based, and Deep Neural Methods for De-Identification of Dutch Medical Records. arXiv preprint arXiv:2001.05714.

[16] 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.

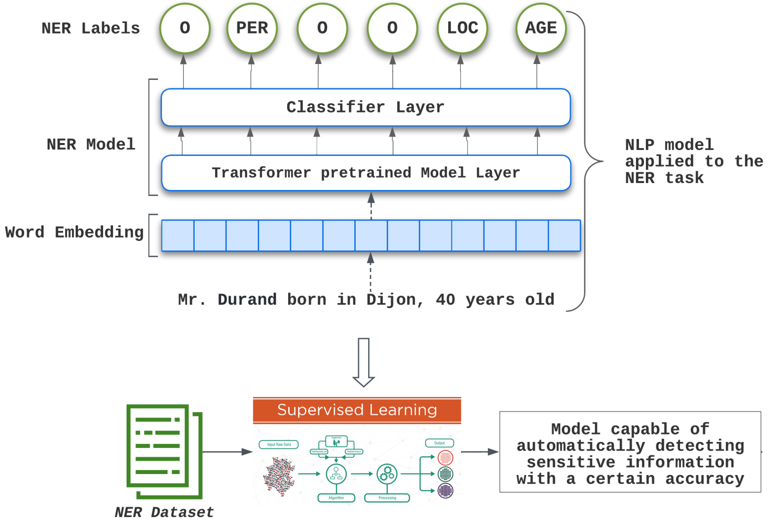
HNFC-NER-TRAIN Labelled Dataset for DL Training

Methodology: 25 hours, @HNFC, 1 person.

1. Input data: 1500 texts (14925 sentences) of deceased persons, first de-identified and then pre-annotated by the previous hybrid approach.
2. Manually annotated @HNFC with Doccanno again.



FLAUBERT NER Model Architecture



NER results

Improved results for almost all metrics

Methods	Hybrid Syst[??]			PROPOSAL			Denoncourt System (RNN)[17]		
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**

- Still not as strong as English-language results.

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

Plan

Applying ϵ-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: represented as a vector, further normalized.

Randomization: Exponential Mechanism

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

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.

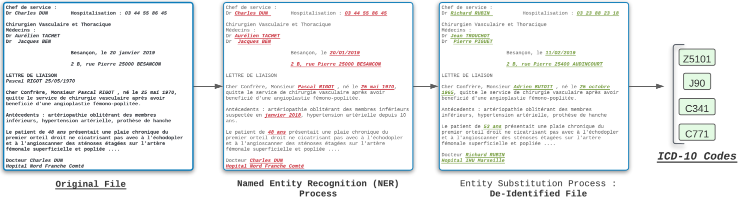
Mr. Julien born in Besançon, 37 years old was admitted to the hospital from 20/02/2020 to March 01, 2020 following a road accident in Besançon.

De-Identification Tool

Plan

ICD-10 Codes

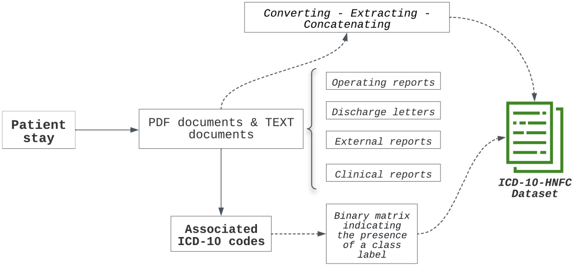
- ICD-10 (International Classification of Diseases, Tenth Revision) codes:
  - A standardized system used for classifying and coding diseases, injuries, and other health-related conditions.
- Assigned to medical diagnoses and procedures to facilitate accurate and consistent recording and reporting of health information.
- Each healthcare stay is manually summarized into ICD-10 codes for statistical purposes and remuneration.
- In the field of computing, it involves 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 consisting of medical texts paired with their respective ICD-10 codes.
- Output: 56,014 very long lines with concatenated results and their corresponding binary vectors of labels.
- Second output: The same text and ICD-10 codes grouped by families, which involves 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

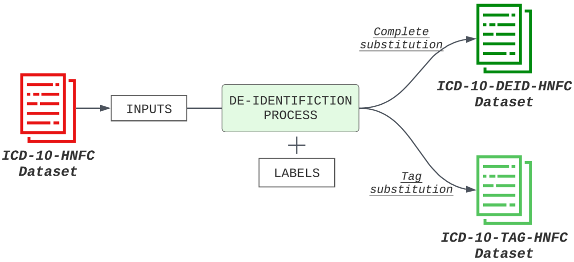
Two issues in ICD-10 codes association

1. Input patient file: Typically a long sequence.
   - Average sequence length is 747, which exceeds the maximum input size for Transformers (512), posing a scalability issue.
2. Large number of different codes and labels, but with sparsity.
   - There are 6,160 unique ICD codes, out of which 3,722 appear less than 10 times, highlighting scalability and sparsity issues.

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

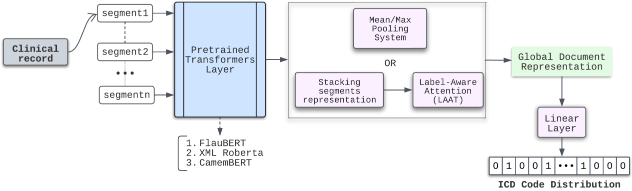


ICD-10 codes association model

Approach with FLAUBERT

- Long sequence processing: Hierarchical Transformers[18].
  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)[19].
  - Labels are integrated into the document representation.

Model Architecture

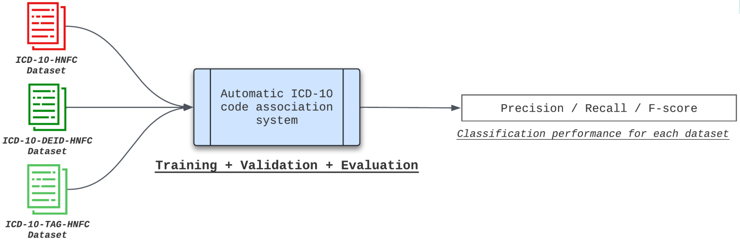


[18] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019, December). Hierarchical transformers for long document classification. In 2019 IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 838-844). IEEE.

[19] 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		0.43	0.41	0.42
ICD-10-DEID-HNFC	6160	0.44	0.43	0.44
ICD-10-HNFC		0.47	0.46	0.47

- ICD-10-DEID-HNFC: Enabled us to prototype the entire ML approach.
- ICD-10-DEID-HNFC vs. ICD-10-TAG-HNFC: Most accurate, close to the original ones.

State of the art of ICD-10 codes association

Experimental results

Models	Language	Dataset	Labels	F_1-score
PLM-ICD[20]	English	MIMIC2	5,031	0.5
		MIMIC3	8,922	**0.59**
Bouzille[21]	French	own dataset	6,116	0.39
			1,549	0.52
		ICD-10-HNFC	6,161	0.27
			1,564	0.35
PROPOSAL			6,161	**0.45**
			1,564	**0.55**

- Bouzille: Uses the same parameters as those in[21]
- All codes (Bouzille and ours) will be on GitHub very soon.
- State-of-the-art ICD-10 codes association model[22] in French language.

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

[21] 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.

[22] 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 the French language.
- Substitution method that combines **utility** and **safety**.
  - Not location-specific Method.
  - Solution available on GitHub[23].

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 the French language.

Future work

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

[23] Surrogate Generation in De-identification, 2022