De-identification of medical reports for machine learning tasks: application to ICD-10 code association

J.-F. Couchot[1], D. Laiymani[1], A. Rahmani[2], P. Selles[2], Y. Tchouka[1]

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

[2]Hôpital Nord Franche-Comté (HNFC), France

Nov. 26th. 2024

Plan

Motivation

De-identification of medical reports for associating ICD-10 code

Leveraging AI for Healthcare Process Improvement

- Significant interest among hospitals in optimizing a number of tasks by leveraging AI, particularly by exploiting **textual medical records** from patient files:
  - Identifying similarities between patients and their pathologies, thereby **gaining direct access to successful treatment pathways** for these pathologies versus less successful ones.
  - **Detecting abnormal patient journeys**, for example, where a condition associated with treatment is suspected.
  - **Automatically associating medical codes** with patient journeys (according to the ICD-10 classif.) for statistical purposes and hospital reimbursement. (@ HNFC $\approx$ 12 individuals from the Medical Information department carry out this coding task e.g.)

Bridging the Gap: Developing AI Tools for Healthcare with sanitized Data

Who can can develop ML to for Medical Institutions?

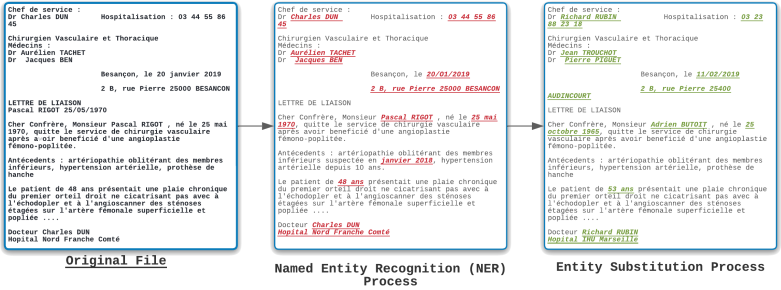
- Medical institutions generally **lack the expertise** to develop advanced AI (with diverse data types like text, tables, and vectors).
- Even within these institutions, **legal restrictions** (GDPR. . .) **prevent AI researchers** (FEMTO-ST) from accessing patient data to develop AI tools.

Is de-identification the answer?

- AI tool prototypes can be created in labs and then **customized with realistic data derived from a robust and useful de-identified medical** data from a medical institution.
  - Robust: the level of information **leakage is mathematically bounded**: based on **Differential Privacy**, the standard adopted today in academia, in industry
  - Useful: surrogating preserves **chronology of events**, distinguish between **personal and medical details** (e.g., "Charcot"), and maintain **familial relationships**.

De-Identification: A Twofold Method

Two Steps



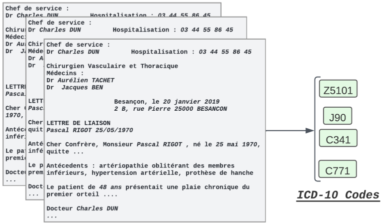
1. Named Entity Recognition (NER) for identifying information (efficiency issue)
2. Sanitizing of detected information (optimization issue: minimizing leakage while preserving utility)

Application Context With ICD-10 Code Association Task

ICD-10: Standardized Diagnostic Tool for Recording Health Conditions

- Developed by the World Health Organization (WHO).
- Used worlwide to classify diseases, injuries, and health conditions. . .
- Reimbursement Impact: Codes are essential for billing and reimbursement systems.

Application Context with ICD-10 Code Association Task



- Manual Coding: Currently, healthcare professionals assign codes manually based on medical records.
- Automated Coding: This task can be framed as a multi-label text classification problem, where the goal is to automatically assign appropriate ICD-10 codes to medical documents.

PhD Thesis of Dr. Yakini TCHOUKA

- Defended in December 2023.
- Supported by the Bourgogne-Franche-Comté region and the EUR EIPHI.
- In partnership with the Medical Information Department of the HNFC.

Plan

Motivation

De-identification of medical reports for associating ICD-10 code

Named Entity Recognition for HIPAA Categories

Iterative Learning on HNFC Datasets: HNFC-NER-EVAL, HNFC-NER-TRAIN

- **Increasingly** large, **progressively** more de-identified datasets.
- Automatically pre-labeled and manually validated.
- Model: Hybrid[1], then deep learning only[2].

NER Results

Method	CamemBERT-ner			MEDINA			FlauBERT-ner			**Hybride**			**Healthinf**			**Dernoncourt**		
Dataset	*HNFC*															i2b2		
Metric	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PER	89	99	93.8	***98.2***	97.7	*98.2*	91.8	97.6	94.6	96.3	***99.8***	98	97.2	98.9	98	**98.2**	99.1	**98.6**
ORG	7.	21.8	11.1	32.6	24.8	28.1	16.9	34.1	22.6	41.1	*57.3*	47.8	*90*	51	*65.6*	**92.9**	**71.4**	**80.7**
LOC	46	67.2	54.6	98.8	81.1	89.1	75.7	66.3	70.7	88.4	***95.8***	92	***99.4***	94.4	***96.9***	95.9	95.7	95.8
DATE	NA			97.7	86.6	91.9	NA			97.7	86.7	91.9	***99.2***	*95.7*	*97.4*	99	**99.5**	**99.2**
AGE	NA			91.5	66.9	77.3	NA			91.5	66.9	77.3	*98.2*	*91.8*	*95*	**98.9**	**97.6**	**98.2**
TEL	NA			99.5	97.9	98.7	NA			***99.5***	97.9	98.7	99.4	***99.8***	***99.6***	98.7	99.7	99.2
REF	NA			NA			NA			NA			96.1	79.5	87	NA		
QID	NA			NA			NA			NA			77.2	32	45.3	**99.2**	**98.7**	**99**
Mic.-avg.	70.8	51.5	59.6	98.2	91.2	94.5	85.8	86.7	86.3	94.6	94.9	94.7	***98.5***	*96.4*	*97.4*	98.3	**98.5**	**98.4**

[1] Tchouka, Couchot, Coulmeau, et al. 2022, "De-Identification of French Unstructured Clinical Notes for Machine Learning Tasks".

[2] Tchouka, Couchot, and Laiymani 2023, "An Easy-to-Use and Robust Approach for the Differentially Private De-Identification of Clinical Textual Documents".

Surrogate generation strategies: DATEs and AGEs

Temporal data surrogate issues

1. **Privacy**:
   - Very identifying
   - Re-identification risk: the chronology of events
2. **Utility**:
   - The relevance of events
   - The patient's features

Related Work on Date Substitution: Uniform Shifting of DATEs

- MIMIC3[3], I2B2[4] datasets.

Attack on HNFC-NER-EVAL Dates with Uniform Shifting

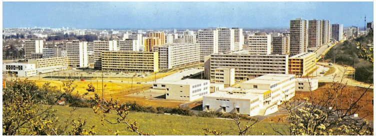
- The interval $I = [I_1, \dots I_{n-2}]$ is NOT modified and is unique in 98% of this dataset.

[3] 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.

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

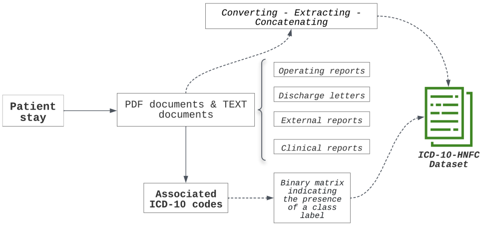
Sanitizing Integrating Metric-Privacy

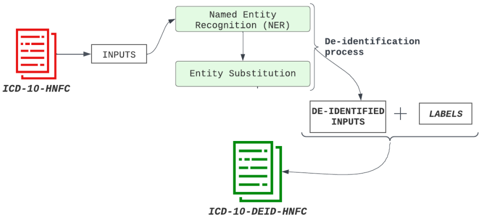
- Theory: $\forall x_1, x_2, y, \Pr(\mathcal{M}(x_1) = y) \leq e^{\varepsilon . d(x_1, x_2)} \Pr(\mathcal{M}(x_2) = y)$.
- Dates: $\mathcal{M}_{date}(x) = x + v$ t.q. $v \sim Lap(\frac{1}{\varepsilon})$.
  - Allows to distinguish betw. 08/01/42 and 14/03/18 (birth and death dates of St. Hawking) whereas DP not.
- Locations: $\Pr(\mathcal{M}_{loc}(x) = o) \propto e^{\varepsilon . d(x,o)}$, s.t. d an epidemiological based distance.
  - Avoid to sanitize Dijon with Beze too often (in BFC but epidemiologically $\neq$).



ICD-10 Code Association[5]

Datasets Buildings





[5] Tchouka, Couchot, Laiymani, Selles, et al. 2023, "Automatic ICD-10 Code Association: A Challenging Task on French Clinical Texts".

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 Code Association– Results

State-of-the-Art[6] Code Association Results

Models	Language	Dataset	Labels	F_1-score
PLM-ICD[7]	*English*	*MIMIC 2*	*5,031*	*0.5*
		MIMIC 3	*8,922*	***0.59***
Dalloux[8]	*French*	*Personnel*	*6,116*	*0.39*
			1,549	*0.52*
PROPOSAL	French	ICD-10-HNFC	6,160	**0.47**
			1,564	**0.55**
Dalloux			6,160	0.27
			1,564	0.35

Impact of De-identification on Results

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

[6] Tchouka, Couchot, Laiymani, Selles, et al. 2024, “Differentially private de-identifying textual medical document is compliant with challenging NLP analyses: Example of privacy-preserving ICD-10 code association”.

[7] Huang, Tsai, and Chen 2022, “PLM-ICD: automatic ICD coding with pretrained language models”.

[8] Dalloux et al. 2020, “Supervised Learning for the ICD-10 Coding of French Clinical Narratives”.

GitHub Open source implementation

- Automatic ICD-10 code classification system in French[9]
- Surrogate generation strategies in de-identification with metric privacy mechanism[10]
- Named Entity Recognition system in medical context[11]
- Automatic ICD-10 code association with CNN[12]

[9] Automatic ICD-10 code classification system in French. https://github.com/mlfiab/icd10-french
[10] Surrogate Generation in De-identification. https://github.com/mlfiab/surrogate-deid
[11] Named Entity Recognition system in medical context. https://github.com/mlfiab/ner-french
[12] Automatic ICD-10 code association with CNN. https://github.com/mlfiab/cnn-icd10