De-identification of Spanish healthcare free-text: not fully reliable but far better than nothing!

Sabrina L. López, Luciano Silvi, Laura Alonso Alemany and Laura Ación

ARPHAI — Gestión epidemiológica basada en inteligencia artificial y ciencia de datos

instituto de cálculo — UBA - CONICET

CONICET

FAMAF UNC

THE PROBLEM

Electronic Health Record (EHR) is valuable data for **secondary use** research, public health policies, etc...

BUT!

it contains free-text with **personal information** that can allow **patient identification**.

Se visita domicilio Av. Belgrano 742. Asisten dr. Perez Natalia MP1234 y agente sanitario Roberto Carlos.

paciente (DNI 1234567) dolores con problm de sustancias. csv. Afebril.

osteooprosis y fragilidad osea
sosp sme de Bruck
Parto Natural 01/01/21

WHY THIS IS IMPORTANT

- Privacy is a human right
- Surveillance, discrimination

DIFFICULTIES

Text

1. Grammatical phrases
2. Typos
3. Ambiguities
4. Frequent use of acronyms

Information

What can lead to the identification of a patient?

Models

- Limited tools in Spanish
- Trained with clinical cases artificially enriched with personal information

Our approach

Developing a *pipeline* to deidentify free-text based in regular expressions and dictionaries.

- Open
- Readable
- Low computational cost
- Locally designed
- Adaptable

Case study: EHR of La Rioja (Argentinian province) from Primary Healthcare Centers.

- Records: 2.394.499
- Period: from 2016-10-04 to 2021-01-28
- Patients: 214.308
- Over 400 records 7.75%!

Results

MANUAL

Annotation guidelines

- 22 entities
- 2 annotators

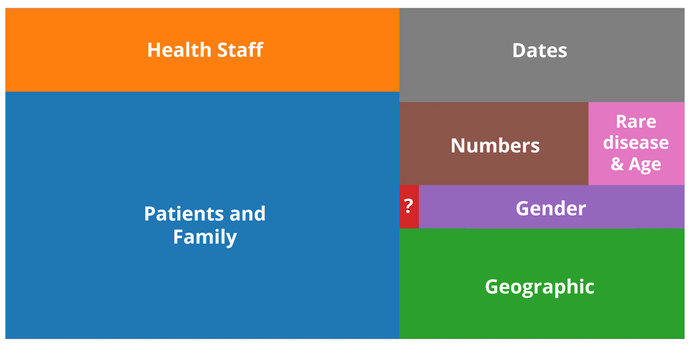
Annotated Dataset

- 2500 EHR
- 57% with entities

Public consultation

for updating the National Personal Data Law

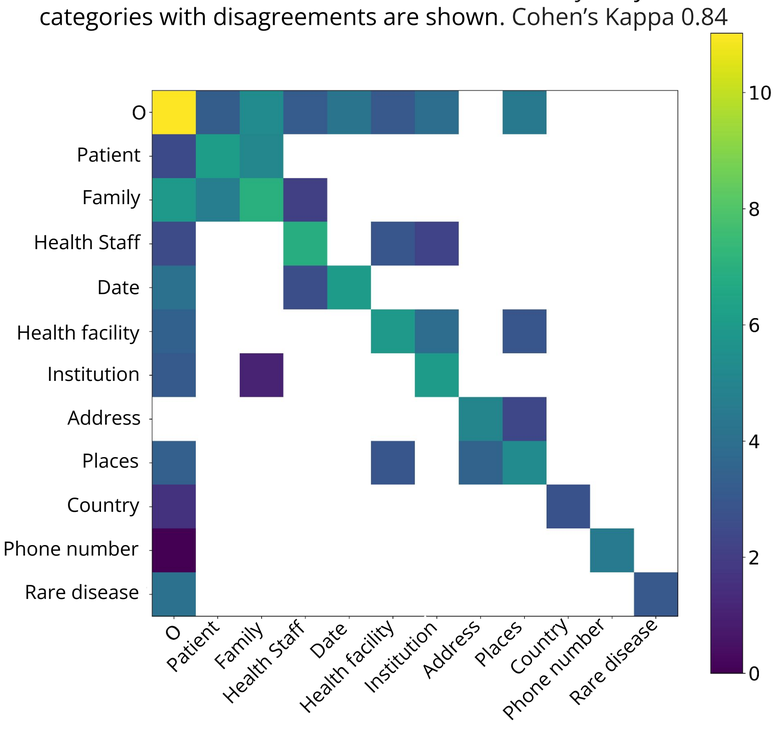
Categories of the 6.111 entities manually detected in 1.442 over 2500 EHR. For clarity, entities are grouped.



Try it yourself!

"no pudo viajar a Bolivia. Juan se enojó, Pedro quiere convivir"

Confusion matrix between annotators. For clarity, only identified categories with disagreements are shown. Cohen's Kappa 0.84



AUTOMATIC

De-identification algorithm

- python
- 17 entities
- Open data dictionaries

Evaluation metrics for three strategies of de-identification

Strategy	Recall	Precision	F1
1*- spaCy + regular expressions without data access	0.11	0.29	0.14
2- regular expressions + dictionaries with data access	**0.41**	0.62	0.44
3- Vanilla biLSTM CRF	0.0	0.03	0.01

*by Instituciones Abiertas

Strategy 1*

Se visita <DIRECCIÓN>. Asisten <DRX> MP1234 y agente sanitario <DRX>.
paciente (DNI 1234567) dolores con problm de sustancias. csv. Afebril.
osteooprosis y fragilidad osea
sosp sme de <PERSONA>
Parto Natural 01/01/21

Strategy 2

Se visita domicilio Av <PERSONA> 742 Asisten dr <DRX> MP1234 y agente sanitario <DRX>
paciente DNI <NUM_DNI> <PERSONA> con problm de sustancias csv. Afebril.
osteooprosis y fragilidad osea
sosp sme de Bruck
<PERSONA> Natural <FECHA>

TAKEAWAYS

- you cannot guarantee that all sensitive information will be removed from EHR free text, by any means
- humans disagree
- automatic approaches do not outperform humans
- better than nothing to mitigate the ever more frequent data breaches!

RECOMMENDATIONS

- limit (secondary) usages of data to the minimum
- anonymize to minimize information leakage

CONTACT US!

sabrina.lopez.ds@gmail.com