UAB
Universitat Autònoma de Barcelona

Desarrollo y validación de *sero*: Una herramienta para la anonimización automatizada de documentos clínicos

Máster de Investigación Clínica
Aplicada a Ciencias de la Salud (ICACS)

Directores
Javier Bracchiglione Xavier Bonfill

Tutor
Xavier Bonfill

Alumno
Ramón José Opazo Campusano

5 de mayo de 2025

1. Introducción | Motivación

- Los datos clínicos *no estructurados* son una fuente de información de alto valor potencial:
  - Riqueza informativa no capturada en fuentes estructuradas
  - Soporte a la toma de decisiones clínicas
  - Investigación en salud
  - Mejora de la gestión sanitaria
- La explotación de estos datos presenta importantes desafíos:
  - Presencia de datos sensibles
  - Heterogeneidad lingüística y contextual
  - Calidad y ruido en los datos
  - Falta de estándares comunes
  - Barreras legales y técnicas

1. Introducción | Otras herramientas

- Ejemplos:
  - Philter
  - ARX Data Anonymization Tool
  - MITRE MIST (foco en la anonimización de documentos industriales)
  - PhysioNet DeID (foco en la anonimización de datos de pacientes hospitalizados)
  - …
- Limitaciones:
  - Orientados a normativa HIPAA
  - Orientados a soluciones industriales
  - Sin trazabilidad integrada
  - Coste elevado de implementación
  - Flujo de trabajo complejo

1. Introducción | Objetivos

- Objetivo general:
  - Desarrollar y validar una herramienta de anonimización automatizada de documentos clínicos que facilite la explotación datos no estructurados
- Objetivos específicos:
  - Desarrollar el software
  - Validar la anonimización de los datos sensibles
  - Validar la securitización de la herramienta
  - Recoger feedback de usuarios

2. Método | Desarrollo del software

- Requisitos funcionales:
  - Anonimización irreversible, *estructural* y *semántica,* de archivos PDF
  - Trazabilidad de documentos a fuente original
  - Securitización robusta de la trazabilidad
  - Ergonomía (facilidad de uso)
- Entorno técnico:
  - Tipo de implementación: web application
  - Lenguaje de programación: Python, TypeScript
  - Framework: FastAPI, PyMuPDF, Astro, DuckDB
  - Requisitos de sistema o plataforma: poder correr Python 3.13 o superior

2. Método | Corpus

- Corpus: conjunto *representativo* de documentos clínicos reales a anonimizar
- Criterios de elegibilidad:
  - Fuente clínica
  - Presencia de datos sensibles que requieran anonimización
  - Redactado en castellano y catalán
  - Antigüedad inferior a 12 meses al momento de realizar la validación
  - Formato PDF con texto reconocible, no imagen escaneada
- Incluye:
  - 50 informes de anatomía patológica
  - 20 epicrisis de pacientes hospitalizados
- Generación del corpus será realizada por técnicos documentalistas del Servicio de Epidemiología Clínica del Hospital Sant Pau

2. Método | Validación de la anonimización

- Se realizará la anonimización a todos los documentos del corpus
- Cada documento se anonimizará usando tres estrategias:
  - Anonimización estructural
  - Anonimización semántica
  - Anonimización estructural y semántica
- Análisis de la anonimización:
  - *Error global < 10%:*
    Los técnicos documentalistas contarán todos los documentos en los que aparezca un caracter o más perteneciente a dato sensible que debió haber sido anonimizado según la estrategia adoptada
  - *Error crítico < 1%:*
    Los técnicos documentalistas contarán todos los documentos en los que, habiendo al menos un caracter no anonimizado, permita la identificación del paciente o personal sanitario, u otro dato sensible

2. Método | Validación de la securitización

- Se entregará la herramienta a un experto en informática para que realice pruebas de análisis de la seguridad y encriptación de los datos
- Las pruebas se realizarán con 10 documentos *no reales* creados para tal motivo
- El experto debe contar con las siguientes cualidades:
  - Ingeniero informático o similar
  - Al menos 10 años de experiencia de trabajo con bases de datos y API RESTful
  - Conocimiento sobre técnicas de ataque a datos encriptados
- El informe técnico debe indicar si la securitización es:
  - Robusta: 0 instancias de fallo (definido como el desciframiento de los datos anonimizados)
  - Frágil: ≥ 1 instancia de fallo

2. Método | Feedback de usuarios

- Se realizará una entrevista semisestructurada a los técnicos documentalistas evaluadores de la herramienta para conocer:
  - Ergonomía de la herramienta (facilidad de uso)
  - Dificultades observadas durante el proceso
  - Evaluación global de la herramienta
  - Oportunidades de mejora

3. Conclusiones

- El desarrollo y validación de *sero* permitirá contar con una herramienta de anonimización de datos sensibles en salud
- Facilitación de la explotación de datos en documentos clínicos no estructurados
    - Investigación en salud
    - Gestión sanitaria

Referencias

Agencia Española de Protección de Datos. 2021. *Guía y herramienta básica de anonimización.* Madrid: AEPD. https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/guia-y-herramienta-basica-de-anonimizacion.

Council of Europe. 2019. *Recommendation CM/Rec(2019)2 of the Committee of Ministers to Member States on the Protection of Health-Related Data.* Strasbourg: Council of Europe. https://rm.coe.int/09000016809339f6.

European Commission. 2022. *New Pseudonymisation Tool Will Foster Research and Better Management of Health Data.* https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/new-pseudonymisation-tool...

ENISA (European Union Agency for Cybersecurity). 2021. *Taking Care of Health Data.* https://www.enisa.europa.eu/news/enisa-news/taking-care-of-health-data.

Tannier, Xavier, Perceval Wajsbürt, Alice Calliger, Basile Dura, Alexandre Mouchet, Martin Hilka, y Romain Bey. 2023. *Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse.* arXiv preprint arXiv:2303.13451. https://arxiv.org/abs/2303.13451.

Tamò-Larrieux, Aurelia. 2023. *From Privacy-Enhancing to Health Data Utilisation: The Traces of Anonymisation and Pseudonymisation in EU Data Protection Law.* Digital Society 2:17. https://doi.org/10.1007/s44206-023-00043-5.

European Federation of Pharmaceutical Industries and Associations (EFPIA). 2021. *Data Anonymisation: Balancing Privacy and Progress.* https://www.efpia.eu/news-events/the-efpia-view/blog-articles/data-anonymisation-balancing-privacy-and-progress/.