Buenos Aires, 30 de septiembre 2022

Sra. Directora
Beatriz de Anchorena
Agencia de Acceso a la Información Pública
S / D

De nuestra mayor consideración:

Luego de participar del espacio de diálogo convocado el día 13 de septiembre de 2022 en relación a la actualización de la Ley de protección de Datos Personales, acercamos algunos comentarios en base a la experiencia ganada en el proyecto ARPHAI (*Argentinean Public Health Research on Data Science and Artificial Intelligence for Epidemic Prevention*) los últimos dos años de trabajo con datos reales de Salud de la República Argentina. De esta forma, esperamos complementar, con nuestros aprendizajes interdisciplinares, el trabajo de especialistas de las áreas socio-productivas, académicas, jurídicas y de la seguridad informática presentes en esa reunión o posteriormente convocados, en torno a algunas de las cuestiones propuestas por el anteproyecto de Ley revisado.

El proyecto ARPHAI es un consorcio de investigación argentino liderado por el CIECTI (Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación) y compuesto también por áreas del Ministerio de Salud de la Nación y la Subsecretaría de Políticas en Ciencia, Tecnología e Innovación del Ministerio de Ciencia y Tecnología. Tiene por objetivo desarrollar herramientas piloto de inteligencia artificial y ciencia de datos para ampliar la funcionalidad de historias clínicas electrónicas para apoyar la gestión de futuras epidemias; así como también generar conocimientos y aprendizajes, mediante sinergias entre la gestión e investigación, que permitan nutrir y mejorar las políticas públicas. Cuenta con la participación de investigadore/as de la UBA, la UNER, UNL, UNQ, UNS, UNC, UNICEN, Flacso -entre otros- y está cofinanciado por el Centro Internacional de Investigaciones para el Desarrollo (IDCR) de Canadá y la Agencia Sueca de Cooperación Internacional para el Desarrollo (Sida) de Suecia.

Desde esa perspectiva, compartimos los siguientes resultados de nuestra experiencia de investigación, en particular del trabajo de la línea de *Uso Responsable de Datos* de ARPHAI, que pueden contribuir a mejorar la futura Ley.

1. Durante el proyecto, trabajando con datos reales en procesos activos para anonimizar textos libres de registros de historia clínica electrónica en español de una provincia de la Argentina (es decir, identificar y ocultar datos personales que permiten identificar al titular de los datos en estos registros), **se evidenció** que:

> ***Un equipo de personas formadas y con experiencia en medicina y en la tarea específica de anonimización no pueden acordar una definición precisa sobre qué información de los textos en la historia clínica permite la identificación del titular de los datos.***

Esta imposibilidad ya había sido identificada en el área de procesamiento automático de textos médicos[1], y en el marco del proyecto ARPHAI verificamos que tampoco es posible llegar a un consenso en los textos originados en la Argentina. Existe acuerdo en que informaciones como el número de DNI o el nombre completo son identificadores, pero no hay acuerdo sobre la capacidad identificatoria de otras informaciones, como domicilio, descripción de la estructura familiar, descripción física o de otras características particulares del titular.

En un esfuerzo de identificación manual de estas informaciones en el texto libre de registros de historias clínicas de La Rioja, tres expertos con formación médica de grado y entrenados específicamente para la tarea y el tipo de texto en particular no pudieron alcanzar un acuerdo total con respecto a qué información debería ser anonimizada para garantizar la no identificación del titular de los datos. Más concretamente, el grado de acuerdo entre expertos anotadores humanos guiados mediante un manual de anotación resultó en un coeficiente Kappa $\kappa=0.8$[2], lo cual indica que existe un desacuerdo entre los jueces humanos mayor al esperable simplemente por azar. Este resultado también pone en evidencia las limitaciones de los métodos automáticos basados en datos etiquetados manualmente, como los métodos de aprendizaje automático.

El derecho a la privacidad de los datos sensibles, como son los datos de salud, impide confiar en un método que no ofrece una garantía total de la preservación del derecho. En nuestro estudio sobre el texto libre de registros electrónicos de atención primaria del subsistema público de la salud en La Rioja se encontró una prevalencia de aproximadamente el 8% de información personal que permite la identificación del titular de los datos, lo cual muestra que el margen de error inherente a la indefinición del problema de anonimización afectaría a un gran número de personas.

> ***Es posible realizar investigación con datos de salud a partir de un conjunto mínimo de datos al que se pueden integrar sucesivas capas de datos de acuerdo a necesidades específicas de cada proyecto.***

---

[1] https://hai.stanford.edu/news/de-identifying-medical-patient-data-doesnt-protect-our-privacy, Yoo J, Thaler A, Sweeney L, Zang J. Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data. Technology Science. 2018100901. October 08, 2018. https://techscience.org/a/2018100901/.

[2] Estadístico que representa el grado de acuerdo entre jueces (expertos del dominio que revisaron del texto libre) respecto de las coincidencias que se podrían producir simplemente por azar. Se trata de una métrica estándar en reproducibilidad de resultados y se calcula según la propuesta descrita en: Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement 20 (1): 37–46. doi:10.1177/001316446002000104

Esto contribuye a la protección de datos que podrían ser personales y/o sensibles y que no son indispensables para el cumplimiento de los objetivos del proyecto. Este criterio, a la vez, coadyuva a cumplir con principios precautorios que se proponen evitar futuras situaciones que, si bien en el presente no conllevan riesgos, podrían hacerlo en otros escenarios. Asimismo, estas prácticas de investigación pueden ser complementadas con medidas de seguridad informática en los ámbitos en los que se resguardan y trabajan los datos.

2. Ante los aprendizajes mencionados, las **recomendaciones** que proponemos son:

- **Que los procesos de anonimización (automáticos o manuales) no se tomen como garantía del derecho a la privacidad** en la Argentina por sus limitaciones previamente mencionadas, especialmente en el caso de los datos sensibles de salud.
- **Que el acceso a datos de salud se limite, incluso para investigación, a la mínima cantidad de información requerida o necesaria para alcanzar estrictamente los objetivos** de un proyecto; y que sean provistos en condiciones que contribuyan a conservar la garantía de privacidad (infraestructura segura, acuerdo de confidencialidad, -de ser pertinente- la aprobación de comité de ética del proyecto de investigación solicitante, aplicación de medidas de anonimización, capacitación de los agentes en contacto con los datos, entre otras.)

Este documento fue elaborado por las Dras. Laura Alonso Alemany, Sabrina López y Verónica Xhardez.

Fernando Porta
Coordinador Académico – CIECTI
Investigador Principal – ARPHAI