



PROGRAMA DE INNOVACIÓN  
TECNOLÓGICA EN SALUD PÚBLICA  
Instituto de Salud Pública. Facultad de Medicina. UBA

# Anonimización de datos de salud



# Características de los datos en salud

- + Compartir datos de salud puede ser muy beneficioso para evitar sesgos de recolección, sesgo del investigador y generar descubrimientos por fuera de los ensayos clínicos
- La filtración de datos personales en salud puede ser muy perjudicial tanto a nivel individual como a nivel general



# Ley 25.326 – Protección de datos personales

- Tipos de datos
- Datos personales: Información de cualquier tipo referida a personas físicas o de existencia ideal determinadas o determinables.
- **Datos sensibles: Datos personales que revelan origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical e información referente a la salud o a la vida sexual.**
- **ARTICULO 7°** — (Categoría de datos).
  1. Ninguna persona puede ser obligada a proporcionar datos sensibles.
  2. Los datos sensibles sólo pueden ser recolectados y objeto de tratamiento cuando medien razones de interés general autorizadas por ley. También podrán ser tratados con finalidades estadísticas o científicas cuando no puedan ser identificados sus titulares.
  3. Queda prohibida la formación de archivos, bancos o registros que almacenen información que directa o indirectamente revele datos sensibles. Sin perjuicio de ello, la Iglesia Católica, las asociaciones religiosas y las organizaciones políticas y sindicales podrán llevar un registro de sus miembros.
- **ARTICULO 8°** — (Datos relativos a la salud).
  - Los establecimientos sanitarios públicos o privados y los profesionales vinculados a las ciencias de la salud pueden recolectar y tratar los datos personales relativos a la salud física o mental de los pacientes que acudan a los mismos o que estén o hubieren estado bajo tratamiento de aquéllos, respetando los principios del secreto profesional.
- **ARTICULO 9°** — (Seguridad de los datos).
  1. El responsable o usuario del archivo de datos debe adoptar las medidas técnicas y organizativas que resulten necesarias para garantizar la seguridad y confidencialidad de los datos personales, de modo de evitar su adulteración, pérdida, consulta o tratamiento no autorizado, y que permitan detectar desviaciones, intencionales o no, de información, ya sea que los riesgos provengan de la acción humana o del medio técnico utilizado.
  2. Queda prohibido registrar datos personales en archivos, registros o bancos que no reúnan condiciones técnicas de integridad y seguridad



PROGRAMA DE INNOVACIÓN  
TECNOLÓGICA EN SALUD PÚBLICA  
Instituto de Salud Pública Facultad de Medicina, UBA

# Ley 26.529 - Derechos del Paciente en su Relación con los Profesionales e Instituciones de la Salud.

Art 2 .

c) **Intimidad.** Toda actividad médico - asistencial tendiente a obtener, clasificar, utilizar, administrar, custodiar y transmitir información y documentación clínica del paciente debe observar el estricto respeto por la dignidad humana y la autonomía de la voluntad, así como el debido **resguardo de la intimidad del mismo y la confidencialidad de sus datos sensibles**, sin perjuicio de las previsiones contenidas en la Ley N° 25.326



# Anonimización

- La anonimización es un proceso que depende principalmente de dos factores:
  - La estructura de los datos en cuestión:
    - Cuán identificables son nuestros registros
  - El contexto en el que se comparten esos datos:
    - Qué otras bases de datos disponibles hay
    - Qué capacidad técnica hay
    - Qué interés tiene la información para los demás
    - Qué implicancias tiene perder la anonimización de esa información en particular



# Tipos de variables

- Identificadores directos: Nombre y Apellido, DNI, Fecha de Nacimiento
  - Permiten identificar a alguien casi unívocamente
- Cuasi identificadores: Localidad, Edad, Sexo, Nivel educativo, Situación laboral, etc
  - Combinados estos datos pueden ser utilizados para la identificación de personas. Mientras más única sea la combinación de variables, más riesgo hay de reidentificación.
- Atributos sensibles: Diagnóstico médico, Procedimiento, Salario, Orientación política, etc
  - Generalmente es la información que presenta interés para otras partes. Muchas veces a mayor diversidad, el valor decrece.



# Medición de riesgo

- Conceptos de evaluación:

- K-anonymity: Por cada combinación de variables cuasi identificadoras, cuántas tienen menos de k registros
- L-diversity: Por cada combinación de variables cuasi identificadoras, cuántas tienen menos de L cantidad de categorías de variables sensibles.
- T-closeness: La distribución de la variable sensible en cada combinación de variables cuasi identificadoras debe ser t parecida a la distribución en el total general

- Software:

- ARX
- BIBLIOTECA sdcMicro en R
  - Ambos dan cuenta de cuántos registros tienen riesgo de poder ser identificados una vez suprimidos los identificadores directos





# Estrategias de anonimización

- Supresión de variables (ej. no publicar Localidad)
- Supresión de datos en particular (ej. no publicar la edad de alguien con muchos años)
- Agregación (ej. convertir edad en una variable de intervalos cada 5 años)
- Muestreo (ej. aclarar que de una base determinada sólo se muestran el 50% de los casos)
- Aleatorización de datos (ej. mezclar los datos de edad y sexo)
- Introducción de "ruido" (ej. sumarle o restarle al año de nacimiento un factor aleatorio conocido o desconocido)





# Base de egresos inicial

- Sin datos de fecha de nacimiento, localidad y DNI, la base sigue teniendo un riesgo de más del 90% de casos identificables
- Definir cuál es la información de mayor valor para la publicación (sexo, edad, hospital, diagnóstico por ejemplo) y realizar la evaluación sólo con esas variables
- Introducir alguna estrategia de anonimización para esos datos en particular



# Base de egresos modificada

- Variables cuasi identificadoras: Sexo, Edad en intervalos de 5 años, Código de hospital de internación, Trimestre y Año de ingreso, Trimestre y Año de egreso
- Variables sensibles: Código de diagnóstico, Código de procedimiento, Días de internación
- Variables suprimidas: Fecha de nacimiento, ID, Educación, Ocupación, Localidad, Especialidad, todas las relativas a gestación y parto, Segundo diagnóstico o procedimiento



~/GCBA/anonimización/prueba en r - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

```
Source
Console Terminal
~/GCBA/anonimización/prueba en r/

> print(sdc)
Infos on 2/3-Anonymity:

Number of observations violating
- 2-anonymity: 72 (0.048%)
- 3-anonymity: 1223 (0.818%)
- 5-anonymity: 2585 (1.730%)

-----

> print(sdc, "risk")
Risk measures:

Number of observations with higher risk than the main part of the data: 5493
Expected number of re-identifications: 2829.38 (1.89 %)
> get.sdcMicroObj(sdc, type = "risk")$suda

Dis suda scores table:
-----
Interval Number of records
1 == 0 147841
2 (0.0, 0.1] 1354
3 (0.1, 0.2] 141
4 (0.2, 0.3] 47
5 (0.3, 0.4] 35
6 (0.4, 0.5] 23
7 (0.5, 0.6] 7
8 (0.6, 0.7] 4
9 > 0.7 10
-----

Attribute contribution:
-----
variable contribution
1 CAPITAL 2.431454086
2 SEXO 17.271476224
3 COD_HOS 41.804406683
4 EDAD5 52.392562787
5 MESI4 30.318759068
6 MESE4 29.260466995
7 ANIOI 70.315468266
8 ANIOE 0.002243729
-----
```

Es posible reducir el  
riesgo  
sustancialmente  
pero es necesario  
evaluar el valor de  
la base con la  
estructura  
resultante



# Procesos de anonimización de Historia Clínica

- Algoritmos que buscan y remueven o alteran nombres, fechas, localidades, DNIs, números de teléfono, nacionalidad, IDs, correo electrónico, direcciones
- Algoritmos que buscan términos médicos y el resto lo dejan como asterisco (text scrubbers)
- Algoritmos de clasificación de frases con información personal (PHI) o no

*En general todo está desarrollado en inglés*

---- Otro tema es el de la encriptación





# HIPAA (Ley de Responsabilidad y Transferibilidad de Seguros Médicos de EEUU)

- Exige excluir de la información de los pacientes:
  - Nombres
  - Toda subdivisión geográfica menor a Estado o los primeros tres dígitos del código postal si la unidad es mayor a 20,000 habitantes
  - Todas las fechas salvo año
  - Todas las edades arriba de 89 años (se pueden agrupar en 90 y más)
  - Números de teléfono, Correos electrónicos, Número de identificación, Número de historia clínica, etc
  - Identificadores biométricos y fotos de rostro
  - Cualquier otro identificador que pueda ser utilizado para reconocimiento

