Pseudonimización de información clínica para uso secundario. Aplicación en un caso práctico ISO/EN 13606

R. Somolinos Cristóbal[1], A. Muñoz Carrero[1], M. E. Hernando Pérez[2], M. Pascual Carrasco[1], R. Sánchez de Madariaga[1], O. Moreno Gil[1], J.A. Fragua Méndez[3], F. López Rodríguez[3], C. H. Salvador[1]

[1] Unidad de Investigación en Telemedicina y e-Salud, Instituto de Salud Carlos III, Madrid, España, {rsomolinos,adolfo.munoz,mario.pascual,ricardo.sanchez,omoreno,chsalvador}@isciii.es

[2] Grupo de Bioingeniería y Telemedicina, Universidad Politécnica de Madrid, Madrid, España, elena@gbt.tfo.upm.es

[3] Laboratorio de Bioingeniería y Telemedicina, Hospital Universitario Puerta de Hierro Majadahonda, Majadahonda, Madrid, España, {jafragua,flopez}@idiphim.org

Resumen

Este trabajo describe la importancia de la pseudonimización de la información clínica para usos secundarios, los aspectos legales que regulan la transferencia de dicha información y las principales técnicas y modelos de anonimización existentes. Se propone el uso de un sistema pseudonimizador conforme a la norma ISO/EN 13606, diseñado y desarrollado por nuestro grupo de trabajo, para la anonimización sistemática de conjuntos de extractos de historia clínica electrónica. Finalmente se muestra la utilización de este servicio de pseudonimización en un proyecto de investigación concreto junto al estudio poblacional previo necesario para establecer los parámetros más adecuados para la anonimización.

1. Introducción

El uso secundario de información clínica para fines docentes, estadísticos y de investigación está en continuo crecimiento en los últimos años. La disponibilidad de datos clínicos públicos para uso secundario es fundamental para el avance en el conocimiento médico.

Un requisito básico para la existencia de repositorios públicos de datos clínicos es garantizar la privacidad de los datos del paciente mediante técnicas de anonimización y de-identificación. Cada país, a través de sus leyes, protege la privacidad de los datos estableciendo diferentes restricciones para el uso secundario de los datos clínicos. De forma general, el intercambio de información clínica para uso secundario sólo es permitido si la información ha sido previamente anonimizada para evitar futuras asociaciones entre los datos y sus propietarios.

De acuerdo a la norma ISO/TS 25237 "Health informatics ï Pseudonymizationò [1], la anonimización es el proceso que elimina todos los enlaces entre un conjunto de datos y el propietario de los mismos, y la pseudonimización es un tipo particular de anonimización que elimina las asociaciones entre los datos y sus propietarios y añade nuevas asociaciones entre un conjunto de datos y uno o más pseudónimos.

Muchas investigaciones necesitan conocer, además de los datos clínicos, ciertos datos demográficos de los pacientes para poder extraer resultados significativos. Si los datos que pueden identificar a una persona (sexo, edad, etc) no son eliminados y se mantienen junto a la información clínica, entonces existe un riesgo de re-identificación de los propietarios de los datos. Según un estudio de Sweeney [2], el 87% de la población de Estados Unidos puede ser identificada unívocamente mediante sólo tres datos: sexo, fecha de nacimiento y código postal.

Nuestro grupo de investigación ha desarrollado en los últimos años un sistema pseudonimizador conforme a la norma ISO/EN 13606 [3]. Este sistema permite realizar pseudonimizaciones sobre conjuntos de extractos ISO/EN 13606 de Historia Clínica Electrónica (HCE). El servicio permite mantener algunos datos demográficos (sexo, fecha de nacimiento y lugar de residencia) junto a la información clínica. Los usuarios de esta herramienta pueden seleccionar entre diversas opciones de granularidad para cada uno de los tres datos demográficos. De esta forma, los investigadores pueden elegir la configuración más adecuada para su proyecto, estableciendo un equilibrio entre los datos demográficos presentados para poder obtener resultados y el riesgo de re-identificación.

2. Aspectos legales

El auge del uso secundario de la información clínica ha provocado que diferentes gobiernos y organizaciones hayan tenido que actualizar sus legislaciones en los últimos años para adaptarse al nuevo entorno en temas relacionados con el acceso y uso de datos personales.

La Unión Europea, a través de su directiva 95/46/EC y el artículo 29 de su grupo de trabajo, ha establecido los mecanismos necesarios para garantizar la protección del individuo en cuanto al manejo y libre circulación de datos personales entre sus estados miembros. Esta directiva de protección de datos no es de aplicación cuando el individuo no puede ser identificado, directa o indirectamente. El artículo 29 establece que los datos anónimos son toda aquella información relacionada con una persona que no puede ser identificada. De acuerdo a la Organización Mundial de la Salud (OMS), existe ñanonimidad razonable o proporcionalò cuando no se puede identificar a los individuos a través de medios razonables. En 2012, la Comisión Europea propuso una

reforma integral de las leyes sobre protección de datos de 1995 para fortalecer los derechos de privacidad y fomentar la economía digital europea.

Las leyes españolas siguen la directiva europea 95/46/EC. De acuerdo a la ley española 14/2007 de investigación biomédica, (artículo 50, 2) los datos de una persona sólo podrán ser usados para propósitos de investigación o docencia si la parte interesada ha dado expresamente su consentimiento o si dichos datos han sido previamente anonimizados, y (artículo 52, 3) dichos datos sólo pueden ser almacenados con fines de investigación en un formato anonimizado.

En Estados Unidos, la ñHealth Insurance Portability and Accountability Actò de 1996 (HIPAA) se encarga de proteger la privacidad de la información clínica y establecer regulaciones que garanticen la seguridad de la HCE. No existe obligación legal de obtener el consentimiento del paciente para guardar su información clínica siempre que los datos hayan sido previamente de-identificados. La ley de privacidad de la HIPAA se encarga de la protección contra descubrimientos de identidad y proporciona definiciones y normas para la de-identificación de datos clínicos. La ñSafe Harborò de HIPAA define un conjunto de 18 datos denominados ñProtected Health Informationò (PHI), los cuales deben ser eliminados para que los datos clínicos se consideren de-identificados.

Por consiguiente, en todos aquellos escenarios en los que un sistema de información desee enviar información clínica al exterior para uso secundario es necesario que dicha información sea previamente anonimizada.

3. Anonimización

La anonimización de información sensible es un problema ampliamente abordado. En la actualidad ya existe gran cantidad y variedad de soluciones que la implementan. Las soluciones más destacadas son las basadas en técnicas de búsqueda de patrones y de aprendizaje automático de máquinas [4]. Las técnicas de búsqueda de patrones utilizan patrones, reglas y diccionarios para localizar datos que puedan provocar la re-identificación. Las técnicas de aprendizaje automático clasifican palabras en datos clave mediante métodos como máquinas de vectores, árboles de decisión, entropía máxima y campos condicionales. Estas últimas técnicas necesitan un entrenamiento largo y supervisado con un gran conjunto de datos, mientras que las técnicas basadas en búsqueda de patrones apenas necesitan entrenamiento y son fácil y rápidamente modificables cambiando sus reglas y diccionarios. La gran desventaja de los métodos de búsqueda de patrones es que necesitan desarrollar muchos algoritmos complejos por cada dato clave y estos algoritmos son personalizados para cada conjunto de datos, no son generalizables para otros documentos. La ventaja de los métodos de aprendizaje automático es que, una vez realizado el entrenamiento, aprenden rápidamente para reconocer patrones de datos complejos y se adaptan mejor a diferentes tipos de documentos.

Además de estas técnicas de anonimización, también existen varios modelos basados en clustering (k-anonimity, l-diversity y t-closeness) que representan el riesgo de re-identificación de registros previamente de-identificados. Estos modelos tratan de cuantificar el riesgo de que agentes externos puedan obtener información privada a partir de los datos anonimizados ofrecidos para uso secundario y bases de datos de acceso público (informaciones censales). Los atributos de los registros para uso secundario se clasifican en los siguientes grupos según su naturaleza:

- atributos clave: son campos que identifican unívocamente a una persona (nombre, dirección, teléfono, números de identificación). Estos atributos son eliminados en los registros anonimizados.
- cuasi-identificadores: son variables del entorno que por sí mismas no identifican a una entidad unívocamente, pero que junto con otros cuasi-identificadores pueden ser utilizadas para re-identificar a una persona (fecha de nacimiento, sexo, código postal). La anonimización puede eliminar estos datos, mantenerlos o generalizarlos.
- atributos sensibles: es información sensible de las entidades que no debe poder ser enlazada con su propietario (tener una determinada enfermedad). Estos datos son de gran utilidad para usos secundarios y se mantienen en los registros anonimizados.

Las clases de equivalencia se definen como los conjuntos de registros que poseen los mismos valores para un conjunto de cuasi-identificadores seleccionado.

3.1. k-anonimity

La k-anonimity previene descubrimientos de identidad. Este modelo previene que, a partir de un conjunto de cuasi-identificadores, se pueda descubrir la identidad o los atributos clave de una entidad. La k-anonimity no se centra en los atributos sensibles de los registros.

Se define ók ô como el mínimo tamaño de todas las clases de equivalencia establecidas en la anonimización. En otras palabras, para cualquier registro, siempre existen al menos otros ók ô-1 registros con idénticos valores de sus cuasi-identificadores. Por tanto, la probabilidad de re-identificar a una entidad a partir de los valores de un conjunto de cuasi-identificadores es 1/k. Para garantizar un bajo riesgo de re-identificación, se debe garantizar un valor mínimo de ók ô.

El método más habitual para disminuir la probabilidad de re-identificación es la generalización de los cuasi-identificadores. Haciendo los cuasi-identificadores menos específicos aumenta el tamaño de las clases de equivalencia y, por tanto, el valor de ók ô, así se reduce el riesgo de re-identificación. Los registros con valores muy poco usuales deben ser eliminados, ya que aumentan significativamente la probabilidad de re-identificación (una altura de 2,21 m, una edad de 108 años), o agrupados a partir de cierto valor del cuasi-identificador (altura de 2,00 m o más, edad de 80 años o más).

Un problema que presenta este modelo es que si, para los ókô elementos de una clase de equivalencia, todos ellos o un porcentaje muy alto poseen el mismo valor de un atributo sensible, un atacante podrá concluir que cualquier entidad perteneciente a esa clase tiene ese determinado valor del atributo sensible con total certeza o con un porcentaje muy alto de acierto. La k-anonimity proporciona protección contra ataques de descubrimiento de identidad, pero no ante ataques de descubrimiento de atributos sensibles.

3.2. l-diversity

La l-diversity proporciona protección frente ataques de descubrimiento de atributos sensibles. Este modelo mide la variedad de los atributos sensibles. Los atributos sensibles deben ser diversos dentro de cada clase de equivalencia para evitar su descubrimiento.

Se define ólô como el menor número de valores distintos de un atributo sensible dentro de cualquier clase de equivalencia. Significa que siempre, para cualquier registro, existen al menos ólô valores posibles distintos para sus atributos sensibles.

La l-diversity no tiene en cuenta la distribución global de los valores sensibles. Por lo que, aunque en la distribución global un valor de un atributo sensible aparezca sólo un 10%, si existe una clase de equivalencia en la que dicho valor aparece en un 80% de los registros, se está otorgando una información adicional muy importante a los posibles atacantes que pretendan inferir información privada de los registros.

3.3. t-closeness

La t-closeness sí considera la distribución de los valores sensibles. Este modelo mide la similitud entre la distribución de los atributos sensibles en cada clase de equivalencia y la distribución global de todos los registros. La distancia entre la distribución global de un atributo sensible y la distribución de ese mismo atributo en cualquiera de las clases de equivalencia no debe nunca superar un umbral ótô prefijado.

4. Pseudonimizador conforme a la norma ISO/EN 13606

Nuestra unidad ha desarrollado un pseudonimizador basado en la norma ISO/EN 13606 que facilita la realización de anonimizaciones de acuerdo a un conjunto de parámetros de una forma sistemática. El sistema recibe extractos ISO/EN 13606 y, a través de su módulo anonimizador y su servidor demográfico, los anonimiza y devuelve a sus clientes.

El servidor demográfico se encarga del almacenamiento permanente de las entidades demográficas presentes en los extractos y proporciona funciones específicas para la gestión de los identificadores. Mientras que el módulo anonimizador realiza la anonimización propiamente dicha: envía la información demográfica al servidor demográfico y la elimina del extracto, gestiona y sustituye los identificadores presentes en el extracto e incorpora al extracto anonimizado los datos demográficos del sujeto de atención con la especificidad seleccionada para el sexo, la fecha de nacimiento y el lugar de residencia.

El pseudonimizador está diseñado para dar servicio a proyectos de muy diversa índole y que, por tanto, poseen atributos sensibles de muy diferente naturaleza. Los atributos sensibles, dependiendo del caso, pueden ubicarse en localizaciones muy diversas del modelo de referencia de la norma ISO/EN 13606. Por este motivo, resulta muy dificultoso implementar de forma generalista modelos basados en atributos sensibles como l-diversity y t-closeness acordes a la norma ISO/EN 13606. Sin embargo, los cuasi-identificadores sí tienen una clara ubicación dentro del modelo de referencia de la norma: las clases del paquete demográfico. El pseudonimizador implementa el modelo k-anonimity centrándose en los cuasi-identificadores de mayor utilidad para usos secundarios (sexo, fecha de nacimiento y lugar de residencia) y proporciona diferentes posibilidades para la generalización de los cuasi-identificadores seleccionados.

El acceso al servicio de anonimización se realiza mediante un web service a través de la función *anonymizeExtract*. Esta función devuelve el extracto anonimizado y debe ser invocada con los siguientes parámetros de entrada:

- *extract*: es el extracto que se desea anonimizar
- *rootProject*: ñespacio de nombresò que se utiliza para generar todos los nuevos identificadores que aparecen en el extracto anonimizado
- *degreeGender*: grado de especificidad para el cuasi-identificador sexo. Opciones disponibles: a) eliminado, b) incluido
- *degreeBirth*: grado de especificidad para el cuasi-identificador fecha de nacimiento. Opciones disponibles: a) eliminado, b) grupos de 10 años, c) grupos de 5 años, d) año, e) mes, f) día
- *degreeAddress*: grado de especificidad para el cuasi-identificador lugar de residencia. Opciones disponibles: a) eliminado, b) país, c) provincia, d) ciudad, e) código postal, f) todo incluido

5. Ejemplo de pseudonimización

El proyecto CAMAMA (FIS 08/1148) y su continuación CAMAMA2 (FIS 12/01476) son proyectos coordinados llevados a cabo junto al Hospital de Fuenlabrada (Madrid) y el Hospital Clinic (Barcelona). En ellos se estudia el envío automatizado de información clínica entre productores (hospitales) y consumidores (biobancos, registros de casos y otros grupos de investigación). En un principio, su objetivo era cubrir sólo los casos de cáncer, pero finalmente se extendió para abarcar a toda la población de Fuenlabrada. El proyecto pretende alcanzar las 200,000 historias clínicas resumidas, intercambiadas por medio de extractos pseudonimizados, correspondientes a la población total de Fuenlabrada.

Para establecer los grados de especificidad adecuados de los cuasi-identificadores en la pseudonimización de los registros intercambiado en el proyecto ha sido necesario realizar un estudio poblacional previo. Este estudio se ha

basado en los datos de la pirámide de población de Fuenlabrada (Tabla 1) [5].

Edad	Hombres	Mujeres	Total
0-4	5740	5381	11121
5-9	6061	5679	11740
10-14	5676	5240	10916
15-19	5117	4908	10025
20-24	6345	6281	12626
25-29	8474	8052	16526
30-34	9770	9377	19147
35-39	9622	9002	18624
40-44	9099	8821	17920
45-49	8091	8069	16160
50-54	7198	8286	15484
55-59	8094	8275	16369
60-64	5579	4709	10288
65-69	2858	2895	5753
70-74	1428	1670	3098
75-79	1100	1525	2625
80+	1313	2495	3808
	101565	100665	202230

Tabla 1. *Pirámide de población de Fuenlabrada*

Todos los registros del proyecto provienen de ciudadanos de Fuenlabrada, por lo que la información sobre el lugar de residencia no es de utilidad para extraer resultados en usos secundarios. Por este motivo, se decidió no añadir en los registros pseudonimizados datos sobre el lugar de residencia y establecer las clases de equivalencia sin tener en cuenta este cuasi-identificador.

Las clases de equivalencia y el cálculo del valor ók-anô se realizaron utilizando sólo los cuasi-identificadores sexo y fecha de nacimiento. El sexo y la fecha de nacimiento del individuo son de utilidad para inferir conclusiones en usos secundarios. Especificidades inferiores al año en la fecha de nacimiento se consideraron que no aportaban información adicional útil para uso secundario. Por esta razón, sólo se valoraron tres opciones para el cuasi-identificador fecha de nacimiento: año, grupos de 5 años y grupos de 10 años. También hay que indicar que para este cuasi-identificador, en los tres casos, se han juntado todos los registros de 80 o más años en un único grupo.

Combinando las posibilidades de ambos cuasi-identificadores se obtuvieron seis configuraciones posibles. A partir de los datos de la pirámide de población, y suponiendo una distribución uniforme entre los 5 años de cada grupo de la pirámide, se calculó el valor de ók-anô para cada una de las configuraciones. En la Tabla 2 se muestran las seis configuraciones posibles junto a la clase de equivalencia con menor número de elementos de cada una de ellas y el valor de ók-anô

k	Sexo incluido	Sexo eliminado
Año	1100/5=220	2625/5=525
	Hombres 75-79	Total 75-79
5 años	1100	2625
	Hombres 75-79	Total 75-79
10 años	1313	3808
	Hombres 80+	Total 80+

Tabla 2. *Estudio poblacional de Fuenlabrada*

El valor de ók-anô más bajo calculado para las seis configuraciones es 220. Este valor proporciona un riesgo de re-identificación muy bajo (1/220) y pertenece a la configuración que mayor información aporta para uso secundario. Consecuentemente, ésa fue la configuración elegida para realizar la pseudonimización en este proyecto: sexo incluido, fecha de nacimiento año y lugar de residencia eliminado.

6. Conclusiones

El pseudonimizador desarrollado supone una potente y novedosa herramienta para la utilización de datos clínicos ISO/EN 13606 en usos secundarios. Una ampliación del servicio, pendiente para trabajos futuros, consiste en añadir una nueva funcionalidad para calcular automáticamente el valor ók-anô a partir del conjunto de extractos y la configuración de los cuasi-identificadores.

Agradecimientos

Este trabajo ha sido financiado parcialmente por los proyectos CAMAMA (FIS 08/1148), CAMAMA2 (FIS 12/01476) y PITES-ISA (FIS 12/00508 y FIS 12/01305).

Referencias

[1] International Organization for Standardization. ISO/TS 25237:2008 "Health informatics ï Pseudonymizationò.

[2] Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000. http://dataprivacylab.org/projects/identifiability/paper1.pdf (Consultada: Septiembre 2014).

[3] Somolinos R, Muñoz A, Hernando ME, *et al*. EHR Anonymising System Based on the ISO/EN 13606 Norm. *IFMBE Proceedings. XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013 (MEDICON 2013)*, Sevilla, 25-28 septiembre de 2013, pp 1302-1305. (ISBN: 978-3-319-00846-2).

[4] Meystre SM, Friedlin FJ, South BR, *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol*, vol. 10, pp. 70. 2010.

[5] Pirámide de población de Fuenlabrada. Página web del ayuntamiento de Fuenlabrada. http://ayto-fuenlabrada.es/recursos/doc/SC/Estadisticas_y_territorio/36781_1111112013133426.pdf (Consultada: Septiembre 2014).