Guia sobre les tècniques i bones pràctiques de la pseudonimització

TIC Salut Social

Generalitat de Catalunya /Salut

DPD

Contingut

1. Introducció 2
2. Terminologia 3
2.1. Tractament de dades: Pseudonimització vs. Anonimització 4
3. Tècniques de pseudonimització 6
3.1. Comptador 6
3.2. Generador de nombres aleatoris (RNG) 6
3.3. Funció hash criptogràfica 6
3.4. Codi d'autenticació de missatge (MAC) 7
3.5. Criptografia simètrica 7
3.6. Escollint tècniques i polítiques de pseudonimització 7
3.7. Recuperació dels pseudònims 8
4. Pseudonimització a la pràctica: Història Clínica (HC) 9
4.1. HC pseudonimitzat amb tècnica de comptador 11
4.2. HC pseudonimitzat amb generador de nombres aleatoris 11
4.3. HC pseudonimitzat amb funció hash 12
4.4. HC pseudonimitzat amb codi d'autenticació de missatge (MAC) 13
4.5. HC pseudonimitzat amb tècnica de xifratge determinista 14
4.6. Pseudonimització i protecció de dades 14
5. Casos d'ús de pseudonimització en Salut i Recerca 16
5.1. Comparació d'Històries Clíniques 16
5.2. Utilització d'Històries Clíniques amb finalitats de Recerca 17
5.3. Emmagatzematge compartit d'Històries Clíniques 18
6. Conclusions i recomanacions 20
7. Annex (I): Escenaris de pseudonimització 21
8. Annex (II): Tècniques d'atac més comunes 22
8.1. Atac de força bruta 22
8.2. Atac de diccionari 22
8.3. Atac per conjectura 22
9. Annex (III): Eines per pseudonimitzar o anonimitzar conjunts de dades 23

Edita:
J. Oriol Castaño Cid[1]. Oficina del DPD - Fundació TIC Salut Social.

1a edició: Barcelona, novembre de 2020.
2a edició: Barcelona, octubre de 2021.

[1] https://www.linkedin.com/in/oriol-castaño/

1. Introducció

Aquesta guia de bones pràctiques té com objectiu explorar les nocions bàsiques de la pseudonimització explicant les solucions i implementacions tècniques disponibles que es poden aplicar a la pràctica, seguint les recomanacions establertes en l'RGDP[2] i en les publicacions d'ENISA[3, 4] en aquest camp.

En primer lloc s'expliquen en detall les diferents tècniques i polítiques de pseudonimització utilitzades en la actualitat, de cara a conèixer els paràmetres que ens permetran decidir quina tècnica i política utilitzar en cada cas. Seguidament es planteja un escenari fictici on s'apliquen aquestes tècniques de pseudonimització a un conjunt de dades d'històries clínics i s'explica els avantatges i els inconvenients d'utilitzar les diferents tècniques en aquests tipus d' escenaris. A continuació es detallen tres casos d'ús de pseudonimització en Salut i en Recerca: comparació d'històries clíniques, utilització de dades amb finalitats de recerca i emmagatzematge compartit d'històries clíniques. Finalment s'extreuen les conclusions i recomanacions per a totes les parts interessades pertinents al que respecta a l'adopció pràctica i implementació de la pseudonimització de dades.

A l'Annex I s'enumeren els diferents escenaris de pseudonimització, definint els principals actors que intervenen en aquest procés a través dels seus possibles rols, i a l'Annex II s'analitzen les tècniques d'atac més típiques utilitzats contra la pseudonimització amb l'objectiu de reidentificar els pseudònims. Finalment a l'Annex III s'analitzen algunes de les eines disponibles actualment al mercat que poden servir tant per pseudonimitzar com per anonimitzat conjunts de dades.

Com s'explica al llarg d'aquesta guia no hi ha una única solució de pseudonimització que funcioni per a tots els enfocaments i escenaris possibles, per contra es requereix un alt nivell de competència per arribar a aplicar un procés de pseudonimització robust i que permeti reduir les amenaces de reidentificació mantenint el grau d'utilitat necessari per al processament de les dades pseudonimitzades.

Aquest document està adreçat a responsables i encarregats del tractaments de dades, desenvolupadors de productes, serveis i aplicacions, així com delegats i autoritats de protecció de dades.

[2] https://gdpr-info.eu/
[3] https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices
[4] https://www.enisa.europa.eu/publications/data-pseudonymisation-advanced-techniques-and-use-cases

2. Terminologia

En aquest capítol es presenten els termes més utilitzats al llarg de l'informe i que són essencials per a la comprensió del lector:

Dada personal: qualsevol informació sobre una persona física identificada o identificable. Per exemple: un nom, un número d'identificació, dades de localització, diversos elements propis de la identitat física, fisiològica, genètica, psíquica, econòmica, cultural o social d'aquesta persona.

Anonimització[5]: Tècnica que permet fer anònimes unes dades per impedir la identificació personal, de manera que no sigui possible reidentificar-la ni revertir el vincle o la connexió que les unia; són dades, doncs, en què la informació s'ha separat de la persona (considerades dades dissociades o anònimes). En les **dades anonimitzades** s'ha trencat el fil conductor entre una informació i una persona física, per això es parla també de dades dissociades, i no se'ls aplica la normativa de protecció de dades.

Pseudonimització: Tècnica de confidencialitat de les dades personals que consisteix a substituir les informacions que identifiquen una persona per un pseudònim. La **pseudonimització** es configura per tant com una mesura tècnica i s'ha de distingir de l'anonimització, ja que a diferència d'aquesta, a les dades pseudonimitzades els hi és plenament aplicable la normativa de protecció de dades ja que es consideren una subcategoria de les considerades "dades personals". Es considera una mesura tant de seguretat com de protecció de dades i aplicada correctament pot rebaixar les obligacions legals del responsable o encarregat del tractament.

Dades pseudonimitzades: dades personals que es poden atribuir a un interessat mitjançant informació addicional. Aquesta informació addicional ha de constar per separat i ha d'estar subjecta a mesures tècniques i organitzatives destinades a garantir que les dades personals no s'atribueixin a una persona física o identificable (art.4.5 RGDP). Per tant es consideren dades personals i estan sotmeses a l'RGPD i a l'LOPDGDD. Un exemple de dada pseudonimitzada és el codi TIP (targeta d'identificació personal) que s'assigna als agents de policia. En el marc de la recerca, implica una separació tècnica i funcional entre l'equip investigador i els encarregats de realitzar aquesta pseudonimització i de conservar la informació que ens pot permetre la identificació en el cas que sigui necessari.

Dades codificades: a diferència de les dades pseudonimitzades, no s'exigeix que en el procés d'obtenir-les hi hagi una separació tècnica i funcional entre l'equip investigador i els encarregats de realitzar aquesta codificació. Hem de distingir-la perquè en l'àmbit de la recerca és usual que els investigadors treballin amb la dada codificada, però aquesta dada codificada en moltes ocasions no es pot qualificar com a dada pseudonimitzada perquè el codi que permet saber o permet deduir la identitat de la persona està en mans del mateix equip investigador. La diferència entre dada codificada i dada pseudonimitzada és rellevant perquè la pseudonimització del reglament de protecció de dades es planteja com a mesura de seguretat i l'estableix en la disposició addicional 17 com a requeriment que ens permet el tractament de dades quan existeix un interès públic en recerca.

[5] https://formacioooberta.eapc.gencat.cat/espaistematics/gestio-dades/anonimitzacio-i-pseudonimitzacio-de-les-dades.html

Esquema d'identificació: Taula on s'associa cada identificador el seu pseudònim corresponent. Depenent de la funció de pseudonimització la taula de mapatge pot ser la contrasenya de pseudonimització o part d'ell.

Identificador: Valor que identifica de manera única a un element (en el nostre cas dada personal) dins d'un esquema d'identificació.

Pseudònim: Valor associat a un identificador d'un individu o qualsevol altre tipus d'informació personal.

Funció de pseudonimització: Funció que substitueix un identificador per un pseudònim.

Contrasenya de la pseudonimització: Paràmetre opcional de la funció de pseudonimització necessari per calcular-la en el cas de que s'utilitzi.

Funció de recuperació: Funció inversa de pseudonimització que mitjançant la contrasenya por recuperar l'identificador associat a un pseudònim.

Entitat de pseudonimització: Entitat responsable de transformar els identificadors en pseudònims mitjançant la funció de pseudonimització. Pot ser un responsable de tractament, un encarregat de tractament que realitzi la pseudonimització en nom del responsable, un tercer de confiança o inclús un interessat tractant-se les seves pròpies dades (veure Annex I). El rol de l'entitat és rellevant per a la implementació pràctica en un escenari específic. La responsabilitat del procés de pseudonimització correspondrà sempre al responsable del tractament de les dades.

Domini de l'identificador i del pseudònim: dominis dels quals s'extreu l'identificador i el pseudònim, poden ser diferents o els mateixos, finits o infinits.

Adversari: Entitat interna o externa que intentat trencar la pseudonimització mitjançant un atac amb l'objectiu d'enllaçar un o més pseudònim amb el seu titular.

2.1. Tractament de dades: Pseudonimització vs. Anonimització

Seguint amb les definicions de pseudonimització i anonimització descrites en l'apartat anterior, cal tenir en compte que transformar un conjunt de dades personals en informació anònima o pseudonimitzada implica realitzar un tractament sobre aquestes dades personals. **El tractament d'anonimització genera un únic i nou conjunt de dades**, mentre que **el tractament de pseudonimització genera dos nous conjunts de dades**: la informació pseudonimitzada i la informació addicional que permet revertir l'anonimització.

- El tractament que genera les dades anonimitzades sí que és un tractament de dades personals, que es pot considerar compatible amb la finalitat original del tractament de dades personals del qual procedeixen les dades (Dictamen 05/2014 sobre tècniques d'anonimització WP246 apartat 2.2.1. Legitimació del procés d'anonimització).
- El conjunt de dades anonimitzades queda fora de l'àmbit d'aplicació de l'RGPD en la mesura que és possible demostrar objectivament que no hi ha capacitat material per associar les dades anonimitzades a una persona física determinada, directament o indirectament, ja sigui mitjançant l'ús d'altres conjunts de dades, informacions o mesures tècniques i materials que hi pugui haver a disposició de tercers.

Aquesta avaluació ha de tenir en compte els costos, el temps requerit per dur a terme la reidentificació o els mitjans tecnològics necessaris per aconseguir la reversió de l'anonimització, tant els actuals com tenint en compte els avenços tecnològics.

D'altra banda el conjunt de dades pseudonimitzades, i la informació addicional vinculada amb aquest conjunt de dades, estan sota l'àmbit d'aplicació de l'RGPD, així com el tractament que els genera. Per aquest motiu **el conjunt de dades pseudonimitzades ha d'estar protegit per quatre tipus de garanties**:

1. En primer lloc, el mateix tractament de pseudonimització que ha **d'impedir la reidentificació** sense disposar de la informació addicional.

2. En segon lloc, els principis i garanties de l'RGPD que estableixen **limitacions**, entre altres, **a les finalitats**, el període de conservació o la comunicació de les dades pseudonimitzades.

3. En tercer lloc, les **garanties** addicionals que incorpori el tractament de les dades pseudonimitzades **en funció del risc** per als drets i llibertats de les persones físiques.

4. En quart lloc, derivat de l'anterior, les garanties tècniques i organitzatives disposades a l'efecte d'impedir la materialització de **bretxes de dades personals**, tant sobre conjunt pseudonimitzat com de la informació addicional.

En canvi sobre el conjunt de dades anonimitzades, des del punt de vista de l'RGPD, només s'aplica un tipus de garanties: la robustesa del **procés d'anonimització** contra la possible reidentificació. Un cop el conjunt de dades està anonimitzat, desapareix l'obligació d'implementar els altres tres conjunts de garanties, almenys des del punt de vista de la normativa de protecció de dades.

Els drets i llibertats dels interessats han d'estar igualment protegits tant en els tractaments d'anonimització com en els processos de pseudonimització.

Acabem resumint les **motivacions d'utilitzar tècniques de pseudonimització**:

(i) És un requisit de l'RGPD,

(ii) Redueix el risc per als interessats,

(iii) permet als responsables del tractament reduir l'esforç a l'hora d'implementar mesures tècniques i organitzatives,

(iv) les dades pseudonimitzades no permeten als destinataris identificar als interessats excepte mitjançant l'ús d'informació addicional,

(v) i que les mesures de protecció de la informació addicional garanteixin que la identificació no es pot produir.

Risc (dades identificades) ≥ Risc (dades pseudonimitzades) + Risc (informació addicional)

3. Tècniques de pseudonimització

En aquest capítol es defineixen les principals tècniques disponibles per pseudonimitzar un identificador únic, comparant les seves característiques d'implementació i fent referència als principals criteris que l'entitat que dugui a terme la pseudonimització pot utilitzar per seleccionar una tècnica o una altra. L'entitat de pseudonimització pot ser diferent segons l'escenari en el que ens trobem (veure Annex I).

3.1. Comptador

És la funció de pseudonimització més **senzilla**. Els identificadors són substituïts per un número triat per un comptador on, primer, una llavor s s'estableix a 0 (per exemple) i després s'incrementa (1, 2 ,3, etc.). És fonamental que els valors produïts pel comptador no es repeteixin mai per evitar ambigüitats.

Els avantatges es basen en la seva **simplicitat**, que el converteixen en un bon candidat per a conjunts de dades petits i no gaire complexos. Pel que fa a la protecció de dades, el comptador proporciona pseudònims sense relació amb els identificadors inicials (tot i que **el caràcter seqüencial del comptador pot proporcionar informació sobre l'ordre de les dades dins del conjunt**). Aquesta solució, però, pot tenir **problemes d'implementació i escalabilitat** en casos de conjunts de dades grans i més sofisticats, ja que caldria emmagatzemar la taula completa de mapatge de la pseudonimització.

3.2. Generador de nombres aleatoris (RNG)

És un mecanisme que extreu valors d'un conjunt de nombres que tenen la mateixa probabilitat de ser seleccionats i que, per tant, són imprevisibles. Hi ha dues opcions disponibles per aconseguir aquest mecanisme: un generador de números aleatoris reals o un generador pseudoaleatori criptogràfic. Cal tenir en compte que en ambdós casos, sense la deguda cura, es poden produir col·lisions (el risc de col·lisió es pot reduir considerablement si els pseudònims generats son de longitud gran, per exemple de 100 dígits).

Aquest mecanisme **proporciona una protecció de dades sòlida** (ja que, contràriament al comptador, s'utilitza un número aleatori per crear cada pseudònim, per la qual cosa és difícil extreure informació sobre l'identificador inicial). Com s'ha esmentat anteriorment, **les col·lisions poden ser un problema, així com l'escalabilitat** (cal emmagatzemar la taula completa de mapatge de pseudonimització).

3.3. Funció hash criptogràfica

Una funció hash criptogràfica pren entrades de dades de longitud arbitrària (com podrien ser els identificadors originals) i els hi assigna sortides de longitud fixa o hash (serien els identificadors pseudonimitzats). Compleix les següents propietats:

- Unidireccional: deduir l'entrada original a partir del hash és inviable a no ser que s'emprin mètodes de força bruta (anar provant tots els missatges d'entrada).

- Sense col·lisions: No poden existir dues entrades diferents amb el mateix valor de sortida (hash).

Com hem dit, s'aplica directament a l'identificador per obtenir el pseudònim corresponent: $Pseudo = H(Id)$, si bé una funció hash pot contribuir significativament a la integritat de les dades, en general **es considera feble com a tècnica de pseudonimització, ja que és propensa a atacs de força bruta i diccionari.**

Com a tècnica de **pseudonimització més avançada** es podrien tornar a **aplicar funcions hash al conjunts de hash generats per la funció inicial**, com per exemple *h3 = hash (h1, h2)*, per aconseguir pseudònims estructurats que es podrien descobrir només parcialment. D'aquesta manera, una cadena de pseudonimització implicaria **diverses entitats** que prenen els pseudònims creats per l'entitat de pseudonimització anterior i crear-ne de nous. Aquesta cadena es mantindria fins i tot si un adversari aconseguís totes les pseudonimitzacions aplicades a la cadena total excepte una, **fent-la una tècnica molt robusta** (és una pràctica habitual en els assajos clínics).

3.4. Codi d'autenticació de missatge (MAC)

En criptografia, un codi d'autenticació de missatge (MAC, en anglès) és una breu informació emprada per autenticar un missatge, és a dir, per confirmar que el missatge prové de l'emissor que diu ser-ho (autenticitat) i que no ha estat modificat en trànsit (integritat).

S'aplica de forma molt similar a la funció hash criptogràfica, amb la diferència **que s'introdueix una clau d'accés per generar el pseudònim**, necessària també per poder recuperar els identificadors a partir dels pseudònims. Es considera una **tècnica robusta de pseudonimització des del punt de vista de la protecció de dades**, ja que la reversió del pseudònim és inviable (sempre que la clau d'accés no sigui compromesa). En els propers capítols s'expliquen variacions d'aquest mètode amb diferents requisits d'utilitat i escalabilitat per part de l'entitat de pseudonimització.

3.5. Criptografia simètrica

La criptografia simètrica (xifratge determinista) i, en particular, l'esquema de xifratge per blocs com l'AES, es pot utilitzar per pseudonimitzar un identificador mitjançant una clau secreta, que serà tant el secret de pseudonimització com el secret de recuperació. L'ús de xifrats per blocs requereix tenir en compte la mida del bloc d'entrada: la mida dels identificadors pot ser menor o major que la mida del bloc d'entrada del xifratge de blocs, si és menor s'ha d'aplicar un esquema de farciment i si l'identificador és major es pot comprimir per a que sigui inferior a la mida del bloc. El xifratge per blocs determinista es pot considerar una **tècnica de pseudonimització robusta**, amb diverses **propietats similars al codi d'autenticació de missatge**.

3.6. Escollint tècniques i polítiques de pseudonimització

Com hem comentat, l'elecció d'una tècnica i política de pseudonimització depèn de diferents paràmetres: del nivell de protecció de dades que es necessiti assolir així com la utilitat del conjunt de dades un cop pseudonimitzat.

En termes de protecció; el generador de nombres aleatoris (RNG), els codis d'autenticació de missatges (MAC) i el xifratge son les tècniques més fortes ja que dificulten els atacs de força bruta, de diccionaris i de conjectures.

Pel que fa a les polítiques de pseudonimització; la totalment aleatòria ofereix el millor nivell de protecció però impedeix qualsevol comparació entre bases de dades. Les funcions deterministes i aleatòries de documents proporcionen més utilitat, però permeten la vinculació entre registres. Podrien ser aplicables solucions específiques en funció dels identificadors que s'hagin de pseudonimitzar (com podrien ser la adreça IP o el correu, explicats posteriorment).

A més, l'entitat de pseudonimització també pot estar preocupada per la complexitat associada a un determinat esquema en termes d'implementació i escalabilitat, tant per la complexitat d'aplicar pseudonimització als identificadors com per la afectació a la mida de la base de dades un cop incorporats els pseudònims. A la taula següent es fa una comparació de les diferents tècniques en termes de flexibilitat i de la mida del pseudònim generat:

Tècnica de pseudonimització	Mida de l'identificador	Mida pseudònim (m en bits)
Comptador	Qualsevol	$m = log2k$
Generador aleatori	Qualsevol	$m >> log2k$
Funció hash	Qualsevol	Fixe o $m >> log2k$
Codi d'autenticació	Qualsevol	Fixe o $m >> log2k$
Xifratge	Fixe	Fixe o igual que l'Id

La majoria de solucions es poden aplicar a identificadors de mida variable, excepte en el cas del xifratge per bloc. La mida del pseudònim depèn de k, que és el nombre d'identificadors que conté la base de dades. Per al generador de nombres aleatoris, la funció hash i el codi d'autenticació de missatges, hi ha una probabilitat de col·lisió i per tant la mida del pseudònim s'ha de triar acuradament. Les funcions hash i els codis d'autenticació es poden dissenyar de tal manera que es garanteixi que la mida del pseudònim eviti qualsevol risc de col·lisió. Finalment, la mida dels pseudònims produïts per xifratge pot ser fixa o igual l'identificador.

3.7. Recuperació dels pseudònims

El mecanisme de recuperació que l'entitat de pseudonimització ha d'implementar pot ser més o menys complex segons la tècnica que s'hagi utilitzat inicialment. En general el mecanisme consisteix en utilitzar el pseudònim i la contrasenya de pseudonimització per recuperar l'identificador original. Pot ser necessari emprar el mecanisme en el cas de detectar una anomalia al sistema i haver de contactar amb les entitats afectades, també en el cas de violació de seguretat s'hauria de notificar als interessats (segons l'RGDP) o inclús com mètode de recuperació per permetre l'exercici dels drets a les persones interessades (articles 12-21 de l'RGDP). A la següent taula es comparen les diferents tècniques respecte als seus mètodes de recuperació:

Tècnica de pseudonimització	Mètode de recuperació
Comptador	Taula de mapatge
Generador aleatori	Taula de mapatge
Funció hash	Taula de mapatge
Codi d'autenticació	Taula de mapatge
Xifratge	Desxifratge

Podem veure que el xifratge és l'únic mètode que no requereix a l'entitat de pseudonimització mantenir una taula de mapatge entre identificadors i pseudònims podent-se aplicar el desxifratge directament sobre el pseudònim per obtenir-ne l'identificador original.

4. Pseudonimització a la pràctica: Història Clínica (HC)

Utilitzant les tècniques i polítiques de pseudonimització presentades anteriorment, en aquest capítol es mostra un cas pràctic: la pseudonimització dels identificadors d'un registre d'històries clíniques de pacients d'un hospital.

Cada història clínica (HC) està associada a un pacient i s'identifica inequívocament a través del **Codi d'Identificació Personal (CIP)**, format per un conjunt regles expressades amb números i lletres que, de manera individual i unívoca, permeten identificar a cada persona acreditada del CatSalut. El codi CIP és únic i es considera dada personal ja que es tracta d'una informació identificativa sobre una persona física, també es considera dada personal de categoria especial relativa a salut ja que va associada a les dades sociosanitàries que apareixen a l'HC de la persona.

Cal, per tant, protegir els identificadors (codis CIP) dels registres d'històries clíniques emprant proteccions com la pseudonimització evitant així que es vinculin amb individus específics. Dit això és vital escollir una tècnica de pseudonimització adequada per trobar un **equilibri entre protecció de dades i utilitat de la informació**.

El **contingut de la història clínica** es determina en l'Article 10 de la Llei 21/2000 d'autonomia del pacient i inclou els següents grups de dades: a) Dades d'identificació del malalt i de l'assistència, b) Dades clinico-assistencials i c) Dades socials.

A continuació es mostra, a través d'un exemple de registre d'històries clíniques (amb **dades fictícies**), com s'aplicarien les diferents tècniques de pseudonimització explicades anteriorment sobre l'identificador CIP. En concret per aquest exemple s'utilitza el grup de dades d'identificació del ciutadà i d'assistència.

Donem per cas que es necessita proporcionar informació dels pacients ingressats a un hospital a l'empresa encarregada del càtering. **L'hospital, com a responsable del tractament, determina les finalitats de l'encàrrec**: facilitar el conjunt de dades dels pacients (que han de ser les mínimes dades necessàries per a dur a terme la finalitat i han d'estar prèviament pseudonimitzades **aplicant aquelles mesures tècniques i organitzative**s destinades a garantir que les dades personals no s'atribueixin a una persona física o identificable) a l'empresa encarregada de preparar el càtering de manera que pugui efectuar correctament el seu servei a l'hospital. Aquest conjunt de dades ha d'incloure: un identificador únic per cada pacient (pseudonimitzat), la data d'ingrés, la unitat assistencial, el número d'habitació i llit, i el metge responsable.

Per dur a terme la preparació del conjunt de dades i la **pseudonimització**, s'escull com a **encarregat del tractament** per compte de l'Hospital a la **fundació TIC Salut Social**, que **proporcionarà les dades pseudonimitzades** a l'empresa de càtering, i que un cop proporcionat el servei haurà d'informar-ho a l'encarregat de tractament.

En cas que l'hospital necessiti saber a quins pacients se'ls hi ha subministrat el càtering, haurà de demanar a l'**encarregat del tractament** que torni a **identificar les dades pseudonimitzades**.

Dades d'identificació del malalt i de l'assistència									
Codi Identificador Personal	Cognoms, Nom	Data naixement	Sexe	Adreça habitual	Telèfon	Data ingrés	Unitat assistencial	Nº hab - llit	Metge responsable
BIOR0290511032	Biniambres Oria, Manola	11/5/1929	F	Avda. Explanada Barnuevo, 77	6467880164	6/12/2020	Atenció sociosanitària	144 - 1	Dr. Higini Espino
HEVA1320705051	Hernaez Varandalla, Leira	5/7/1932	F	Paseo Junquera, 51	9322403255	26/9/2020	Atenció especialitzada	039 - 1	Dra. Ana Matos
TRTU1350828033	Tramarria Turrilla, Víctor	28/8/1935	M	C/ Fernández de Leceta, 36	6555783916	27/11/2020	Prestacions complementàries	003 - 1	Dr. Eduard Rodriguez
DUBA0360430023	Durante Balanzuela, Cipriano	30/4/1936	M	Rua da Rapina, 31	6450137461	13/12/2020	Atenció especialitzada	063 - 2	Dr. Carlos Novoa
TEBO0441004043	Teijon Bolante, Esperanza	4/10/1944	F	Comandante Izarduy, 85	6111724802	29/8/2020	Atenció farmacèutica	037 - 2	Dr. Carlos Novoa
PELU0470316025	Pedrero Luis, Tasiana	16/3/1947	F	Salzillo, 1	6701441490	7/3/2020	Atenció sociosanitària	163 - 1	Dra. Mª Alejandra de Diego
PEPA1470702041	Perejon Parandones, Antonia	2/7/1947	F	Avda. Los llanos, 13	936589584	8/3/2020	Prestacions complementàries	037 - 1	Dra. Mar Salinas
AMAN0531214022	Amada Ansola, Esteban	14/12/1953	M	C/ Amoladera, 36	6756646229	28/11/2020	Atenció primària	016 - 1	Dra. Mª Alejandra de Diego
CATE1540912025	Carriedo Tellez, Isaac	12/9/1954	M	Avendaño, 59	6103018928	3/10/2020	Atenció telefònica i en línia	157 - 1	Dra. Ana Matos
LACO0570227002	Lara Cothal, Bibiana	27/2/1957	F	Calle Aduana, 54	6298466425	4/4/2020	Atenció telefònica i en línia	044 - 2	Dr. Eduard Rodriguez
VIPA0750714005	Villasante Panguision, Dorotea	14/7/1975	F	C/ Cuesta del Álamo, 4	6644307267	11/6/2020	Atenció continuada i urgent	151 - 1	Dr. Carlos Novoa
BUAL0761204010	Bustelo Almeida, Castor	4/12/1976	M	Eusebio Dávila, 11	6285551950	22/9/2020	Prestacions complementàries	029 - 2	Dra. Ana Matos
LLJO0791125021	Lloreda Jorge, Ingrid	25/11/1979	F	Cruce Casa de Postas, 55	7969999518	17/6/2020	Atenció farmacèutica	138 - 2	Dra. Mar Salinas
SAIB1791211054	Saez Ibarra, Ataulfo	11/12/1979	M	Alvaro Cunqueiro, 78	9897977341	21/3/2020	Atenció continuada i urgent	044 - 2	Dra. Mar Salinas
YAME1810223010	Yanguas Messia, Lourdes	23/2/1981	F	C/ Señores Curas, 8	9550684187	5/10/2020	Atenció especialitzada	187 - 2	Dr. Higini Espino
PESE1850911031	Pelon Ser, Fabritzio	11/9/1985	M	Bouciña, 90	6773388653	7/5/2020	Atenció especialitzada	101 - 1	Dr. Carlos Novoa
EGMA0870916030	Eguilaz Madera, Quintin	16/9/1987	M	Avda. Rio Nalon, 43	6194749463	1/3/2020	Prestacions complementàries	154 - 1	Dra. Ana Matos
VAHO0910801050	Valverde Hontoria, Alain	1/8/1991	M	C/ Eras, 71	6831050991	2/8/2020	Atenció a la salut mental	171 - 1	Dr. Eduard Rodriguez
OMDI1981221011	Ompanera Dios, Davinia	21/12/1998	F	La Fontanilla, 29	6598123328	19/9/2020	Atenció especialitzada	041 - 2	Dra. Mar Salinas
PEOJ1010504005	Pelayo Ojeda, Sergio	4/5/2001	M	C/ Canarias, 30	9317834673	24/9/2020	Atenció telefònica i en línia	162 - 1	Dr. Eduard Rodriguez

*Taula 1. Conjunt de **dades fictícies**, d'identificació del pacient i de l'assistència, d'un registre d'històries clíniques*

4.1. HC pseudonimitzat amb tècnica de comptador

Del conjunt de dades originals, s'eliminen aquelles no necessàries per la finalitat del tractament (noms, dates de naixement, sexe, adreça i telèfon dels pacients). El codi d'identificació personal (CIP) es pseudonimitza i la resta d'informació es manté intacta ja que és necessària per dur a terme el servei de càtering.

La pseudonimització de l'identificador es pot realitzar segons les tècniques explicades en el capítol 3, a continuació es mostra la taula reduïda i pseudonimitzada mitjançant la tècnica més simple, emprant un comptador:

Dades d'identificació del malalt i de l'assistència				
PSEUDÒNIM	**Data ingrés**	**Unitat assistencial**	**Nº habitació - llit**	**Metge responsable**
1	6/12/2020	Atenció sociosanitària	144 - 1	Dr. Higini Espino
2	26/9/2020	Atenció especialitzada	039 - 1	Dra. Ana Matos
3	27/11/2020	Prestacions complementàries	003 - 1	Dr. Eduard Rodriguez
...	...	...	...	...

Taula 2. Tècnica de pseudonimització: Comptador (1, 2, 3, etc)

Aquesta tècnica, com les posteriors, permet recuperar la identitat dels pseudònims amb l'ús d'una taula de mapatge, com la que es mostra a continuació:

Taula relacional	
CIP	**Pseudònim**
BIOR0290511032	1
HEVA1320705051	2
TRTU1350828033	3
...	...

Taula 3. Taula relacional CIP – Pseudònim generat pel comptador

4.2. HC pseudonimitzat amb generador de nombres aleatoris

En aquest cas la pseudonimització es realitza generant un nombre aleatori que formi part d'un univers suficientment gros, per exemple entre l'1 i el 10.000, de manera que s'eviti el risc de que es generi el mateix pseudònim per diferents codis identificadors.

Dades d'identificació del malalt i de l'assistència				
PSEUDÒNIM	**Data ingrés**	**Unitat assistencial**	**Nº habitació - llit**	**Metge responsable**
903	6/12/2020	Atenció sociosanitària	144 - 1	Dr. Higini Espino
6.481	26/9/2020	Atenció especialitzada	039 - 1	Dra. Ana Matos
4.665	27/11/2020	Prestacions complementàries	003 - 1	Dr. Eduard Rodriguez
...	...	...	...	...

Taula 4. Tècnica de pseudonimització: Generador de nombres aleatoris (RNG) De l'1 al 10.000

Taula relacional	
CIP	**Pseudònim**
BIOR0290511032	903
HEVA1320705051	6.481
TRTU1350828033	4.665
DUBA0360430023	9.110
...	...

Taula 5. Taula relacional CIP – Pseudònim generat amb RNG

Tant aquest mètode com l'anterior no revelen cap informació sobre els identificadors inicials i no permeten fer cap anàlisis addicional dels pseudònims. Amb tal d'augmentar la utilitat és possible aplicar la pseudonimització només a una part del codi CIP o afegir-li algun valor característic per permetre anàlisis rellevants. Per exemple el nombre de pacients que provenen de la mateixa unitat assistencial, o pacients que comparteixen metge, planta o habitació.

En funció del nivell de protecció de dades i utilitat que hagi d'aconseguir l'entitat de pseudonimització, poden ser possibles diferents variacions conservant altres parts d'informació en els pseudònims. En aquests casos pot haver un problema d'escalabilitat, especialment si es requereix que s'assigni el mateix pseudònim al mateix codi CIP, ja que l'entitat ha de fer comprovacions creuades a tota la taula cada cop que es vulgui pseudonimitzar una nova entrada.

4.3. HC pseudonimitzat amb funció hash

El nombre total de possibles codis CIP s'estima aproximadament en 611.000 milions, aquest fet fa que siguin fàcils de trobar o d'endevinar, cosa que converteix aquesta la **tècnica de pseudonimització mitjançant hash en feble**, parlant en termes de probabilitats de reidentificació. Un adversari, intern o extern pot tenir fàcilment accés a una llista codis CIP pseudonimitzats per realitzar un atac de diccionari i reidentificar-los sense un esforç considerable a nivell tècnic i de temps.

Malgrat els riscos sobre protecció de dades esmentats, el valors hash poden ser útils alguns casos. Per exemple per a la codificació interna de pacients o com a mecanisme de validació o integritat d'un controlador de dades. La funció hash també es pot utilitzar per pseudonimitzar parts d'una identificador (per exemple, les inicials del cognom pacient o la data de naixement que formen part del CIP), permetent una cerca utilitat en els pseudònims derivats.

Dades d'identificació del malalt i de l'assistència				
PSEUDÒNIM	**Data ingrés**	**Unitat assistencial**	**Nº hab - llit**	**Metge responsable**
e32755c32d20192d4278fb62415bf34ea59de39284377ff5f1b33d80898d45e7	6/12/2020	Atenció sociosanitària	144 - 1	Dr. Higini Espino
af64a741834a9bb1f7b857e67e6524bf100c02d14e20baf079ed9309fd8ff30e	26/9/2020	Atenció especialitzada	039 - 1	Dra. Ana Matos
0a17f681346793d909679fd19bc2375edba516aa81c1405d16c229349fe234ae	27/11/2020	Prestacions complementàries	003 - 1	Dr. Eduard Rodriguez
aa22436825739d5c2f2db953a143ca7546362c1d35339801180a9b1c72ba7f7c	13/12/2020	Atenció farmacèutica	063 - 2	Dr. Carlos Novoa
3403155c4882df36395e33fd15b1947e1859f38796fcb502acf02f94c0068473	29/8/2020	Atenció sociosanitària	037 - 2	Dr. Carlos Novoa

Taula 6. Técnica de pseudonimització: Funció hash criptogràfica mitjançant l'algorisme SHA256

Taula relacional	
CIP	**Pseudònim**
BIOR0290511032	e32755c32d20192d4278fb62415bf34ea59de39284377ff5f1b33d80898d45e7
HEVA1320705051	af64a741834a9bb1f7b857e67e6524bf100c02d14e20baf079ed9309fd8ff30e
TRTU1350828033	0a17f681346793d909679fd19bc2375edba516aa81c1405d16c229349fe234ae
DUBA0360430023	aa22436825739d5c2f2db953a143ca7546362c1d35339801180a9b1c72ba7f7c
TEBO0441004043	3403155c4882df36395e33fd15b1947e1859f38796fcb502acf02f94c0068473
…	…

Taula 7. Taula relacional CIP – Pseudònim generat per la funció hash

Taula relacional

CIP	Pseudònim
BIOR0290511032	BIOR2EFA9A664C80AF295A3ABADFDFA6266BB9DCD42CFE887E38095FF25A5628D873
HEVA1320705051	3207051E24DDB14F024D943995461AE025443ADECFA0DF6AED698AC7A82B87C7DEC906
...	...

Taula 8. Taula relacional CIP – Pseudònim generat per la funció hash mantenint les inicials dels cognoms o la data de naixement dels pacients buscant certa utilitat

En la taula 8 es mostra com **es pot realitzar la pseudonimització** sobre una part del codi CIP, per **mantenir certa utilitzat en el conjunt de dades pseudonimitzat**. Per exemple en el primer pseudònim es mantenen les inicials del cognom de pacient (Biniambres Oria) i en el segon es manté la data de naixement (05 de juliol de 1932). Aquesta taula relacional permetria a l'encarregat del tractament i a l'empresa de càtering **tenir informació sobre l'edat dels pacients o el cognom** d'aquests, per exemple, però mantenint la resta del codi CIP ofuscat.

4.4. HC pseudonimitzat amb codi d'autenticació de missatge (MAC)

En comparació amb el hash, **el codi d'autenticació de missatge proporciona avantatges significatius de protecció de dades (autenticitat i integritat)** sempre que la contrasenya s'emmagatzemi de forma segura. L'entitat de pseudonimització pot utilitzar aquest mètode per restringir l'accés del responsable del tractament o a un tercer proveïdor en els casos en que l'accés als pseudònims sigui suficient per als propòsits particulars, com podria ser la publicitat basada en els interessos de l'usuari, en que els anunciants han d'associar un pseudònim únic per a cada individu sense revelar la seva identitat. En el nostre exemple també seria vàlid, ja que l'empresa de càtering no necessita saber en cap moment la identitat dels pacients.

Com en la tècnica anterior, per tal **d'augmentar la utilitat** també es podria aplicar la pseudonimització MAC per separat a les diferents parts de l'identificador (any de naixement, inicials dels cognoms, etc.) utilitzant la mateixa contrasenya d'autenticació.

Sobre la recuperació dels pseudònims, cal subratllar que fins i tot l'entitat de pseudonimització que té accés la contrasenya, no és capaç de revertir-los. Aquesta inversió només es pot obtenir reproduint els pseudònims de cada adreça coneguda i comparant-la amb la llista d'adreces pseudonimitzades. En el cas que existeixi una taula de mapatge la reidentificació serà trivial, augmentant, això si, els requisits d'emmagatzematge.

Per aquestes raons, aquesta **no és una tècnica de pseudonimització pràctica** en els casos en què el **responsable de tractament hagi de poder associar fàcilment pseudònims** amb els codis d'identificació dels pacients.

Dades d'identificació del malalt i de l'assistència				
PSEUDÒNIM	Data ingrés	Unitat assistencial	Nº hab - llit	Metge responsable
ee25002a27cac50cd4eeeacd3ca6e49096e79750c08f07493b935595359c3c63	6/12/2020	Atenció sociosanitària	144 - 1	Dr. Higini Espino
7769233debe9c41b655176ca5da08dd5a5406625a8040be775f784c15948559b	26/9/2020	Atenció especialitzada	039 - 1	Dra. Ana Matos
bbe3ecb84eff8492c50992eb512030698680d74a5874d0e3aba4085589ced816	27/11/2020	Prestacions complementàries	003 - 1	Dr. Eduard Rodriguez
26f38976585ff824718abd7641b12525cb039ee30fd0ac3a46256ed6bcaa44a1	13/12/2020	Atenció farmacèutica	063 - 2	Dr. Carlos Novoa
4931652139c56a6c990c99cd9d8f780f7ed35d8fbfbdc865c976ac00450717a6	29/8/2020	Atenció sociosanitària	037 - 2	Dr. Carlos Novoa
...	...	...	...	...

Taula 9. Técnica de pseudonimització: Codi d'autenticació de missatge (MAC) mitjançant l'algorisme SHA256 i emprant la contrasenya de signatura: HV2020Salut@

Taula relacional	
CIP	Pseudònim
BIOR0290511032	ee25002a27cac50cd4eeeacd3ca6e49096e79750c08f07493b935595359c3c63
HEVA1320705051	7769233debe9c41b655176ca5da08dd5a5406625a8040be775f784c15948559b
TRTU1350828033	bbe3ecb84eff8492c50992eb512030698680d74a5874d0e3aba4085589ced816
DUBA0360430023	26f38976585ff824718abd7641b12525cb039ee30fd0ac3a46256ed6bcaa44a1
TEBO0441004043	4931652139c56a6c990c99cd9d8f780f7ed35d8fbfbdc865c976ac00450717a6
...	...

Taula 10. Taula relacional CIP – Pseudònim generat mitjançant l'algorisme SHA256 amb clau

4.5. HC pseudonimitzat amb tècnica de xifratge determinista

Una alternativa al mètode anterior és el de xifratge aplicat de manera determinista (mitjançant una contrasenya per produir els pseudònims). És més pràctic ja que la recuperació és possible directament mitjançant el procés de desxifratge. Tot i això la naturalesa **del xifratge no permet treure utilitat de les dades pseudonimitzades**. Encriptar per separat les parts del codi CIP pot ser suficient per pal·liar aquest problema, de manera similar al codi d'autenticació de missatge.

Per exemple, es podria **xifrar per separat l'any de naixement dels pacients**, així tindríem un conjunt de dades que compartirien una part del pseudònim i sabríem que és perquè son pacients que tenen en comú aquesta característica, que només sabria l'entitat pseudonimitzadora. D'aquesta manera **s'augmentaria la utilitat del conjunt de dades pseudonimitzades**.

Dades d'identificació del malalt i de l'assistència				
PSEUDÒNIM	Data ingrés	Unitat assistencial	Nº hab - llit	Metge responsable
z jY6YGdv|sBF?	6/12/2020	Atenció sociosanitària	144 - 1	Dr. Higini Espino
!{qH7Z@dx{wBH>	26/9/2020	Atenció especialitzada	039 - 1	Dra. Ana Matos
-)o\7ZCdy}zBF@	27/11/2020	Prestacions complementàries	003 - 1	Dr. Eduard Rodriguez
|,]H6ZDdu~rBE@	13/12/2020	Atenció farmacèutica	063 - 2	Dr. Carlos Novoa
-{]V6[Beq{vBG@	29/8/2020	Atenció sociosanitària	037 - 2	Dr. Carlos Novoa
...	...	...	...	...

Taula 11. Tècnica de pseudonimització: Xifratge determinista mitjançant protocol AES-256 utilitzant la contrasenya: HV2020Salut@

Taula relacional (NO ES NECESSÀRIA)	
CIP	Pseudònim
BIOR0290511032	Funció inversa AES-256(BIOR0290511032&HV2020Salut@)
HEVA1320705051	Funció inversa AES-256(HEVA1320705051&HV2020Salut@)
...	...

Taula 12. Taula relacional CIP – Pseudònim generat mitjançant l'algorisme de xifratge AES-256 amb clau privada

En aquest cas no cal que l'encarregat del tractament emmagatzemi la taula relacional CIP - Pseudònim, ja que per obtenir els identificadors originals només cal aplicar la funció de desxifratge AES-256(CIP&CLAU) al codi CIP indicant la contrasenya que s'hagi utilitzat per xifrar-lo inicialment.

4.6. Pseudonimització i protecció de dades

El problema principal de pseudonimització d'identificadors de curta llargada son els atacs de força bruta, de diccionari o per discriminació ja que permeten realitzar reidentificacions

completes sense esforços considerables si la funció de pseudonimització no s'escull correctament (veure Annex II).

Tenint en compte el problema esmentat, les funcions hash criptogràfiques són vulnerables en el cas de pseudonimitzar identificadors com el CIP. En el cas, per exemple, d'utilitzar la funció hash SHA-256, un adversari que disposi del pseudònim pot realitzar de manera satisfactòria un atac de diccionari sense utilitzar grans recursos de computació ni temporals.

Per tant, per a la protecció de dades cal escollir altres funcions de pseudonimització, com ara el codi d'autenticació de missatge, xifratge amb contrasenya, o inclús generadors de nombres aleatoris, ja que, un adversari no pot llençar els mateixos atacs perquè aquests mètodes utilitzen una clau secreta (MAC i xifratge) o l'atzar (per el RNG). També es podria utilitzar el comptador, però sent prudent davant les possibles prediccions derivades de la naturalesa seqüencial del mètode.

5. Casos d'ús de pseudonimització en Salut i Recerca

Les històries clíniques dels pacients poden tenir diversos propòsits: per informar els metges durant el procés assistencial, per organitzacions d'assegurances mèdiques per tal de calcular aspectes financers dels tractaments de malalties, o per organitzacions de recerca per tenir dades estadístiques sobre diagnòstics i medicaments. Més enllà d'aquests, evidentment hi ha molts altres grups d'interès per a les dades mèdiques dels pacients.

El problema d'aquests propòsits diferents és que en cada cas només es necessita accés a determinades parts d'un registre de dades mèdiques, però no necessàriament al registre complet. Un metge necessita tenir accés principalment a les dades mèdiques pertinents, però no necessàriament als aspectes financers relacionats amb l'assegurança, i una companyia d'assegurances no hauria de tenir accés a molts detalls sobre el diagnòstic exacte ni la història clínica. Les organitzacions de recerca mèdica només poden accedir a la informació binària sobre si un pacient és tractat amb un determinat medicament o no, potencialment en combinació amb el diagnòstic, però no han d'accedir en cap cas als identificadors de la persona (com el nom, codi CIP, data de naixement, etc.) ni als antecedents mèdics exactes ni a les dades financeres.

En aquests casos d'ús la pseudonimització pot proporcionar protecció a la informació sensible dels pacients contra l'accés, accidental o intencionat, de qualsevol d'aquestes parts. L'acte de pseudonimització ajuda a separar les dades mèdiques de la identitat del pacient, permetent potencialment realitzar investigacions mèdiques sobre dades pseudonimitzades.

A continuació ens centrem en un hipotètic entorn d'intercanvi d'històries clíniques, per il·lustrar una manera en què la pseudonimització pot protegir la privadesa dels pacients alhora que permet el processament de dades mèdiques amb les finalitats descrites anteriorment.

5.1. Comparació d'Històries Clíniques

Suposem que dos hospitals han de comprovar si disposen de la mateixa versió actualitzada de la història clínica d'un pacient determinat. A causa de retards en la digitalització o transmissió de dades des dels hospitals implicats, es pot produir una situació en què no estigui clar si el registre de dades del pacient als diferents servidors d'emmagatzematge és complet i coherent entre les dues entitats i els seus possibles subcontractats TIC . Per tant, es fa necessari executar un protocol que compari els registres de dades específics del pacient.

Sense l'ús de tècniques de pseudonimització, caldria que un hospital enviés tota aquesta informació al segon hospital, per tal de comparar tots els camps de dades de la història clínica del pacient. Òbviament, d'aquesta manera es revelarien totes les dades mèdiques i personals del pacient afectat al segon hospital, independentment de si aquest registre de pacient existeix o no a l'emmagatzematge de dades d'aquest hospital.

Aquest enviament de dades personals es podria evitar fàcilment mitjançant l'ús d'un esquema d'arbre de hash per a la pseudonimització de la història clínica del pacient abans de la comparació. Aquest enfocament requereix que l'hospital remitent pseudonimitzi cada entrada única del registre de dades mèdiques mitjançant una funció de pseudonimització adequada (per exemple una funció hash) i després torni a aplicar la funció hash a les categories d'informació (dades personals, símptomes, diagnòstics, medicaments), per finalment fer un hash dels quatre tipus de categories (veure Fig. 1).

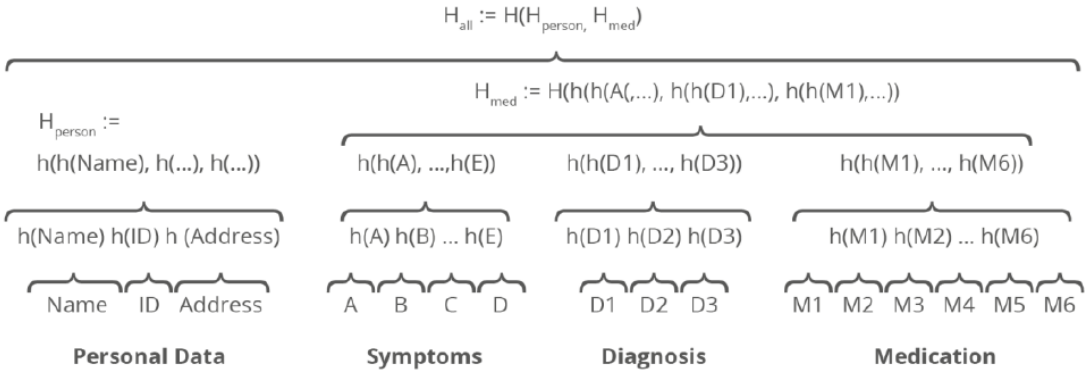

Fig. 1 Esquema d'arbre de hash

Aquests pseudònims no necessàriament representen de manera independent dades personals, com per exemple, el valor hash d'un nom de malaltia per si mateix no està relacionat amb cap individu humà. Tanmateix, en el context d'un registre personal de salut, l'existència o l'absència d'aquest valor de hash evidentment esdevé informació personal del pacient. Per a què aquest esquema funcioni, és de vital importància assegurar-se que símptomes, diagnòstics i medicaments idèntics es representin en un format textualment idèntic.

D'aquesta manera si dos pacients comparteixen el mateix conjunt de símptomes, compartiran alguns pseudònims idèntics: els dels símptomes específics compartits. Això permetria a l'hospital receptor de dades descobrir fàcilment aquests pseudònims i conèixer els veritables símptomes, diagnosis i medicacions d'un pacient.

Seguint l'esquema d'arbre de hash podem arribar a obtenir un pseudònim que representi totes les dades mèdiques d'aquest historial clínic, és a dir, que abasti tots els símptomes, diagnòstics i medicaments. Depenent del nivell de granularitat que es doni en el format específic de registre de dades del pacient, es poden aplicar altres nivells de pseudonimització de manera similar, fins que es creï el pseudònim únic més alt de tots (etiquetat *Hall* a la Fig. 1). Un cop finalitzada aquesta pseudonimització, la tasca de comparació de dos registres de dades de pacients es redueix a la comparació només dels seus pseudònims de nivell superior. Si són idèntics, totes les dades dels dos registres de pacients també són idèntiques. Si difereixen, només el conjunt de pseudònims del següent nivell inferior s'envia al segon hospital, revelant el mínim possible detalls addicionals sobre les condicions mèdiques exactes del pacient.

D'aquesta manera, la comparació de dos registres de dades mèdiques es pot reduir fàcilment a la comparació dels nivells particulars de pseudònims, cosa que permet identificar fàcilment les diferències exactes i els seus elements de dades mitjançant l'execució d'un protocol d'intercanvi de pseudònims adequat.

5.2. Utilització d'Històries Clíniques amb finalitats de Recerca

Més enllà de la simple comparació de registres de dades de pacients, una altra utilització habitual de les històries clíniques és la detecció de correlacions i patrons entre símptomes i medicaments, intentant identificar noves formes de diagnòstic o tractament per a certes malalties. Aquesta tasca no es realitza normalment als mateixos hospitals, sinó que se subcontracta a institucions dedicades a la recerca mèdica que analitzen les dades de moltes fonts diferents.

Per tant, en aquestes institucions s'han d'analitzar les dades de diversos pacients per trobar patrons comuns de rellevància mèdica. Per a aquest tipus d'anàlisi, la identitat dels pacients no és directament rellevant. Una excepció a aquest supòsit es produeix quan les dades del pacient poden revelar un nou diagnòstic, per exemple, a causa de tenir els mateixos patrons de símptomes i medicaments que tots els altres que comparteixen el nou diagnòstic. En aquests casos, és necessari tornar a identificar el pacient en concret per notificar-li (i els seus metges) el nou diagnòstic.

La tasca de detectar correlacions i patrons estadístics en símptomes i medicaments es pot realitzar fàcilment mitjançant la comparació de pseudònims de nivell 1 (Fig. 1) sense ni tan sols revelar el veritable valor del símptoma o medicament subjacent. Per tant, la institució de recerca pot treballar fàcilment només amb pseudònims de nivell 1, sense aprendre mai cap símptoma ni medicació real. D'aquesta manera, la identitat i el registre mèdic personal dels pacients es protegeix sobre manera, tot i que la utilitat prevista de l'anàlisi de dades sobre símptomes i medicaments continua sent factible. Un inconvenient d'aquest enfocament consisteix en la limitació de l'abast de la utilitat: aquest esquema de pseudonimització no admet automàticament altres consultes diferents de la correlació de patrons que es presenta aquí.

Considerem a continuació l'escenari en què s'ha de tornar a identificar un pacient per notificar-li un diagnòstic descobert per la institució de recerca. Òbviament, la institució investigadora no pot ni hauria de poder contactar directament amb el pacient per protegir la seva identitat. Per tant, es fa necessari que la institució investigadora es posi en contacte amb l'hospital d'emmagatzematge de dades per a aquest pacient i provoqui una notificació del pacient realitzada per aquest hospital. Per tractar aquest cas, la institució investigadora només necessita emmagatzemar un identificador de l'hospital on va rebre les dades, el pseudònim personal relacionat amb el pacient (*Hperson* de la Fig. 1), així com el conjunt de pseudònims de nivell 1 rebuts per a aquest pacient d'aquest hospital.

En cas de detectar una afecció mèdica rellevant que requereixi una notificació, aquest pseudònim *Hperson* s'envia a l'hospital del pacient del qual provenien les dades. Aleshores, l'hospital pot reidentificar el pseudònim *Hperson* mitjançant la informació addicional necessària, tornar a identificar el pacient en qüestió i realitzar la actuacions pertinents. D'aquesta manera, la mateixa institució d'investigació de recerca no aprèn mai la identitat del pacient, però pot oferir un nou diagnòstic a aquest pacient.

5.3. Emmagatzematge compartit d'Històries Clíniques

Un altre cas d'ús de la pseudonimització és l'emmagatzematge compartit de dades mèdiques entre diferents entitats, per tal de millorar la seguretat, integritat i disponibilitat de la informació, mitjançant pseudonimització criptogràfica i sistemes de compartició de secrets[6].

El sistema de compartició de secrets que utilitzen algoritmes criptogràfics, com per exemple l'esquema de Shamir[7], és una forma de compartició de secrets on **un secret es divideix en parts** i es dóna a cada participant una sola: **de manera que totes o part d'elles són necessàries per reconstruir el secret.**

[6] https://en.wikipedia.org/wiki/Secret_sharing
[7] https://en.wikipedia.org/wiki/Shamir%27s_Secret_Sharing

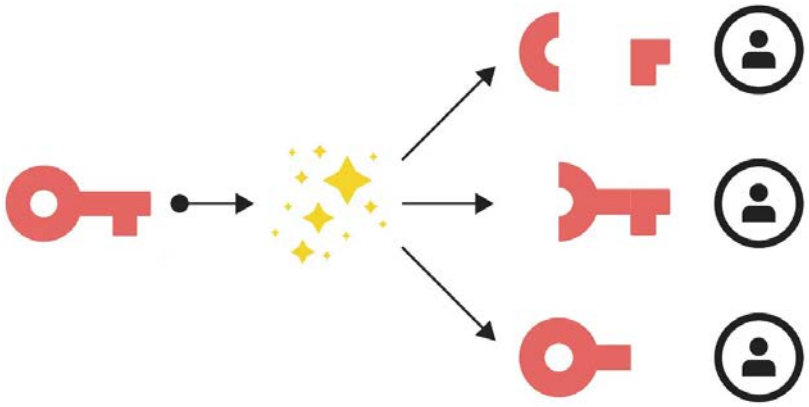

Fig. 2 Secret dividit i compartit en tres parts

En aquest sentit, fins i tot si utilitzem pseudonimització criptogràfica amb clau secreta es pot augmentar el nivell de seguretat mitjançant sistemes de compartició de secrets entre diferents hospitals. Llavors, en cas que sigui necessària una reidentificació del pacient, caldrà que un conjunt suficientment ampli dels hospitals participants, o d'altres entitats participants en el sistema de compartició del secret, proporcionin les seves participacions en conseqüència. Si això passa, es podrà restaurar la clau secreta, i per tant es podrà tornar a identificar el pacient.

Un aspecte interessant d'aquest esquema consisteix en l'enfocament de xifratge adoptat. Si s'utilitza un esquema de **xifratge simètric estàndard**, on el xifratge i el desxifratge utilitzen la mateixa clau secreta, aquesta clau secreta es converteix en el secret per compartir. En aquest cas, ni crear ni descobrir un pseudònim funciona sense accés a aquest secret de pseudonimització. Tanmateix, si s'utilitza un **xifratge asimètric**, hi ha dues claus diferents: clau privada i clau pública. En aquest cas, la clau privada és necessària per reidentificar els pseudònims creats, de manera que cal compartir-la mitjançant un sistema de compartició de secret. La clau pública, però, es pot utilitzar fàcilment per crear nous pseudònims (és a dir, xifrar les noves dades del registre de pacients) sense necessitat de resoldre cap secret compartit. Cada hospital pot decidir afegir dades (personals) a la història clínica d'un pacient, simplement xifrant una addenda al registre original i copiant-la a tots els hospitals que comparteixin la història del pacient en particular. Tot i això, per tal de descobrir la identitat que hi ha darrere de l'addenda xifrada, és necessària la col·laboració d'un conjunt d'hospitals al llarg de les restriccions del sistema de compartició de secrets (per descobrir la clau privada i desxifrar tant els identificadors).

6. Conclusions i recomanacions

Com s'explica durant la guia, la pseudonimització de dades en infraestructures d'informació complexes és un repte, amb una alta dependència del context, les entitats implicades, el tipus de dades, la informació de fons i els detalls d'implementació. De fet, no hi ha una solució senzilla a la pseudonimització que funcioni per a tots els enfocaments en tots els escenaris possibles. Per contra, requereix un alt nivell de competència per aplicar un procés de pseudonimització robust, reduint els possibles atacs de reidentificació i mantenint el grau d'utilitat necessari per al processament de les dades pseudonimitzades.

La pseudonimització d'identificadors pot resultar una tasca difícil i propensa a errors. Tot i això no és l'elecció de la tècnica de pseudonimització la que causa més problemes, sinó la vinculabilitat implícita entre un conjunt de pseudònims i altres valors de dades (com per exemple altres informacions personals d'un pacient en el seu historial clínic). Amb aquestes dades addicionals, encara que els identificadors CIP estigui correctament pseudonimitzats existeix l'amenaça de reidentificació per discriminació.

Com es discuteix en el capítol 3, les tècniques de pseudonimització aleatòria redueixen la vinculació entre diferents pseudònims de diferents conjunts de dades, per tant poden mitigar les possibles estadístiques extretes de les bases de dades pseudonimitzades. Al mateix temps, limiten la possibilitat d'enllaçar diferents registres de dades de diferents conjunts a un sol perfil d'usuari. Seguint els anàlisis fets anteriorment, **el millor enfocament de la pseudonimització és**:

- Considerar tot el conjunt de dades disponibles.
- Conèixer el conjunt dades individuals que s'introdueixin al sistema.
- Aplicar la pseudonimització a totes les dades de manera que els atacs de força bruta i de diccionari siguin inviables.
- Eliminar qualsevol opció d'atacs de distribució estadística o d'informació addicional.
- Dissenyar la funció de pseudonimització resultant de manera que el conjunt de dades pseudonimitzat mantingui només el tipus d'utilitat necessària pel propòsit del tractament, eliminant qualsevol altre utilitat.

A continuació s'extreuen algunes conclusions i recomanacions bàsiques en relació amb l'adopció pràctica i la implementació de la pseudonimització de dades:

Els responsables i encarregats del tractament de dades han de considerar amb deteniment la implementació de la pseudonimització seguint un enfocament basat en el risc, tenint en compte la finalitat i el context general del processament de dades personals, així com els nivells d'utilitat i escalabilitat que volen aconseguir.

Els desenvolupadors de productes, serveis i aplicacions han de proporcionar informació adequada als responsables i encarregats del tractament sobre l'ús de tècniques de pseudonimització i els nivells de seguretat i protecció de dades que proporcionen.

Els reguladors (per exemple, les autoritats de protecció de dades i el Comitè europeu de protecció de dades) han de proporcionar orientacions pràctiques als responsables i encarregats del tractament de dades pel que fa a l'avaluació d'impacte i anàlisis de riscos , tot promovent les millors pràctiques en el camp de la pseudonimització.

7. Annex (I): Escenaris de pseudonimització

A continuació es presenten diferents escenaris de pseudonimització que es poden trobar a la pràctica, enumerant els diferents actors i objectius específics en cada cas.

- **Escenari 1: Pseudonimització per a ús intern**

El responsable del tractament és qui les pseudonimiza per a un posterior processament intern. L'objectiu en aquest escenari és millorar la seguretat de les dades personals, ja sigui per ús intern, per compartir entre diferents unitats o en cas d'incident de seguretat.

- **Escenari 2: Encarregat del tractament involucrat en la pseudonimització**

Aquest escenari és una variació del primer, on també hi participa l'encarregat del tractament de dades obtenint els identificadors dels subjectes en nom del responsable del tractament. No obstant, la pseudonimització la realitza també el responsable, que segueix tenint el rol d'entitat de pseudonimització. L'objectiu és el mateix de l'escenari anterior. Exemple: proveïdor de serveis en el núvol que allotja serveis de recollia de dades en nom del responsable del tractament qui serà l'encarregat d'aplicar posteriorment la pseudonimització a aquestes dades.

- **Escenari 3: Enviament de dades pseudonimitzades a l'encarregat de tractament**

En aquest escenari el responsable del tractament de dades torna a realitzar la pseudonimització, però aquesta vegada l'encarregat del tractament no participa en el procés de recol·lecció de dades, sinó que les rep un cop pseudonimitzades pel responsable (per fer anàlisis estadístics o per emmagatzemar-les), qui també les recopila. D'aquesta manera la pseudonimització protegeix la seguretat de les dades respecte a l'encarregat de tractament.

- **Escenari 4: Encarregat de tractament com a entitat de pseudonimització**

Un altre escenari possible és el cas en què el responsable assigna la tasca de pseudonimització a un encarregat del tractament de dades. Les dades pseudonimitzades s'envien al responsable del tractament, que en aquest cas, només emmagatzema dades tractades.

- **Escenari 5: Tercer com a entitat de pseudonimització**

També es pot donar el cas de voler realitzar la pseudonimització a través d'un tercer (no un encarregat) que posteriorment reenviï les dades al responsable del tractament. Contràriament a l'escenari anterior, el responsable no té accés als identificadors dels interessats ja que el tercer no està sota control d'aquest. D'aquesta manera es millora la seguretat i protecció de dades a nivell del responsable d'acord amb el principi de minimització de dades.

- **Escenari 6: Subjecte com a entitat de pseudonimització**

Aquest és un cas especial on els pseudònims son creats directament pels seus mateixos subjectes (interessats). Cada individu genera el seu propi pseudònim i és amb el que a continuació envia les seves dades. L'objectiu d'aquest tipus de pseudonimització és que el responsable de tractament no aprengui els identificadors dels interessants i que aquests puguin controlar el procés de pseudonimització. S'aplica en els casos en que el responsable no necessiti tenir accés a la informació original i compleix clarament amb el principi de minimització de dades.

8. Annex (II): Tècniques d'atac més comunes

En aquest annex es consideren les diferents tècniques d'atac de reidentificació més típiques que poden afectar a la pseudonimització. L'eficàcia d'aquests atacs depèn de diversos factors: la quantitat d'informació del titular que conté el pseudònim, el coneixement de l'adversari, la mida del domini de l'identificador, la mida del domini del pseudònim i de l'elecció i configuració de la funció de pseudonimització. Aquestes tècniques d'atac es descriuen a continuació:

8.1. Atac de força bruta

Aquesta pràctica està condicionada per la capacitat de l'atacant a l'hora de calcular la funció de pseudonimització (sense contrasenya), o a l'accés a la seva implementació (saber que fa la caixa negra de la funció). Si l'objectiu és aconseguir una reidentificació completa el domini identificador ha de ser finit i relativament petit. Per a cada pseudònim trobat, l'adversari pot intentar recuperar l'identificador original aplicant la funció de pseudonimització a cada valor fins que trobi una coincidència.

Si la mida del domini dels identificadors és infinita l'atac de força bruta es fa inviable, si la mida és molt grossa la complexitat és extremadament difícil però deixa a l'adversari el potencial d'un atac de discriminació. En el cas d'utilitzar una contrasenya de pseudonimització l'atacant no podria calcular la funció de pseudonimització sempre i quan no tingués accés a la implementació de la funció i la contrasenya fos segura: s'ha d'aïllar del conjunt de dades, s'ha d'eliminar de forma segura de qualsevol suport insegur, tenir un registre d'accés i les polítiques de control d'accés han de garantir que només les entitats autoritzades tinguin accés a aquesta contrasenya.

8.2. Atac de diccionari

L'atac de diccionari és una optimització de l'atac de força bruta ja que permet estalviar costos de càlcul. Com l'adversari ha de fer front a una gran quantitat de pseudònims per a dur a terme una reidentificació, precalcula un conjunt de pseudònims i desa el resultat en un diccionari. Cada entrada del diccionari conté un pseudònim amb el seu corresponent identificador, i quan l'adversari necessita identificar un pseudònim fa la cerca al diccionari. Aquesta cerca té un cost previ de processament i emmagatzematge, però la reidentificació només té el cost d'una cerca. Aquests atacs poden funcionar fins i tot per a dominis infinits.

8.3. Atac per conjectura

Els atacs per conjectura o suposicions utilitzen coneixements bàsics (com ara la distribució de probabilitat o altra informació lateral) que l'adversari por tenir. L'adversari no necessàriament ha de tenir accés a la funció de pseudonimització i l'atac es pot fer fins i tot quan el domini dels identificadors és enorme ja que la discriminació és possible mitjançant un anàlisis de freqüència dels pseudònims observats. Depenent dels antecedents o metadades que posseeixi l'adversari i de la quantitat d'informació vinculable que es troba al conjunt de dades pseudonimitzades, aquest tipus d'atac pot conduir a descobrir la identitat d'un, d'uns quants o de tots els pseudònims. Especialment per a conjunts de dades petits o si els pseudònims contenen informació addicional que faciliti la reidentificació, aquests atacs poden ser fàcilment factibles.

9. Annex (III): Eines per pseudonimitzar o anonimitzar conjunts de dades

Implementar i aplicar mecanismes o tècniques de pseudonimització pròpies per tal de reduir possibles identificacions indegudes pot esdevenir una tasca complicada tècnicament i que requereixi molt temps. Per aquest motiu a continuació es mostra un resum de les eines més populars disponibles avui dia al mercat per realitzar pseudonimització o anonimització de conjunts de dades personals. Algunes d'aquestes eines implementen una varietat de conceptes i algorismes que admeten multitud de models[8]: *k-Anonymity, ℓ-Diversity, t-Closeness*, privacitat diferencial, etc.

- Eines proporcionades per l'U.S. National Institute of Standards and Technology (NIST)[9]
- Eines recomanades per la Johns Hopkins University[10]
- ARX Data Anonymization Tool[11]
- Amnesia[12]
- μ-ARGUS[13]
- sdcMicro[14]
- Anonimatron[15]
- Aircloak[16]

Per qualsevol dubte o aclariment addicional podeu adreçar-vos al DPD de Salut:
dpd@ticsalutsocial.cat
https://ticsalutsocial.cat/dpd-salut/
Tel.: 93 553 26 42 (9:00 a 14:00 h.)

[8] https://arx.deidentifier.org/overview/privacy-criteria/
[9] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools
[10] https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/
[11] https://arx.deidentifier.org/
[12] https://amnesia.openaire.eu/
[13] http://neon.vb.cbs.nl/casc/mu.htm
[14] https://cran.r-project.org/package=sdcMicro
[15] https://realrolfje.github.io/anonimatron/
[16] https://aircloak.com/

Salut/ DPD