Kockázatokat rejt az egészségügyi adatok anonimizálása

Dr. Alexin Zoltán, Szegedi Tudományegyetem, Természettudományi és Informatikai Kar, Szoftverfejlesztés Tanszék

Az anonimizálás célja az, hogy a személyes adatokat átalakítsa olyan módon, hogy azok már nem kapcsolathatók össze természetes személyekkel. Az anonimizálás megfelelő validálás nélkül magában hordozhatja azt a kockázatot, hogy az adatokat később mégis természetes személyekhez lehessen kapcsolni. Amikor ez kiderül, rendszerint jóvátehetetlen kár következik be, mert az adatokat már megosztották, eladták, vagy nyilvánosságra hozták. A tisztességes anonimizálás számol a kockázatokkal és minden lehetséges eszközzel védekezik az újraazonosítás ellen. A szerző a magyar népesség-nyilvántartás statisztikai adatainak segítségével tárja fel a jelenlegi magyar egészségügyi anonimizálási gyakorlat gyengeségeit.

The goal of the anonymization is to transform personal data such a way that data cannot be linked to natural persons any longer. Anonymization without appropriate validation may always carry certain risk for re-identification. When this fact comes to light data subjects had already suffered irremediable loss since the "anonymous" data might have already been sold, shared or publicized. Fair anonymization counts with this risk of re-identification and fights against it with all possible means. In this paper the author reveals the weaknesses of the current medical anonymization practice by the help of statistical distribution data obtained from the national population registry.

BEVEZETÉS

Az anonimizálás a görög ἀνωνυμία (anonimia) szóból származik, amelynek a jelentése név nélkül, illetve névtelenül. Tudományos kutatók meg szokták különböztetni az ún. de-identified (közvetlen személyes azonosítókat nem tartalmazó) állományokat, amelyeknél számolnunk kell az újraazonosítás kockázatával, és a valóban anonim állományokat, ahol ez a kockázat elhanyagolhatóan kicsi. A szakirodalomban az anonimizálás szót akkor használják, ha a személyes adatok átalakításának az a célja, hogy egy olyan állományt kapjanak, amely esetében az újra-azonosítás kockázata elhanyagolható.

Orvosi kutatás esetén a kutatási alanyok személyiségi jogait védeni kell. Ez nem csak jogi, hanem egyben morális kötelesség is. A védelem egyik módja az, hogy az adatokat anonimizálják mielőtt átadják a kutatóknak. Ideális esetben ez teljes védelmet nyújt az érintetteknek, hiszen senki sem tudja később azonosítani őket, így a jogaik nem sérülnek. Az anonimizálásnak nélkülözhetetlen és jelentős szerepe van az orvosok kommunikációjában, amikor eseteket vitatnak meg konferenciákon, tudományos folyóiratokban.

A magyar egészségügyi államigazgatási rendszer korlátlan törvényi felhatalmazást kapott olyan, jogszabály szerint anonim adatbázisok használatára, amelyben természetes személyazonosító adatok nem szerepelnek ugyan, de megtalálható bennük a születési dátum, a lakóhely irányítószáma és a nem. Egy hasonló adatállományt az USA-ban 1995-ben már feltörtek, és nem sokkal később olyan jogi szabályozás lépett hatályba, amely jelentősen szigorította a születésre és a lakóhelyre utaló adatok használatát anonim adatállományokban. A szerző ebben a cikkében a magyar népesség-nyilvántartás statisztikai adatainak felhasználásával objektív becslést ad a jelenlegi állítólagos „anonim" adatállományok újra-azonosítási kockázatára.

Előzmények

Kezdetben úgy tűnt, hogy néhány jól meghatározott adat törlésével, esetleg egyszerű átalakításával anonim adatokhoz lehet jutni. Paul Ohm a Texas Egyetem jogász professzora [1] cikkében azonban három olyan esetet ismertetett az Egyesült Államokból, amelyek során anonimnek hitt adatállományokat törtek fel egyszerű módszerekkel. 1995-ben Latanya Sweeny végzős egyetemi hallgató sikerrel azonosította a GIC (Group Insurance Commission) egészségbiztosító-társaság „anonim" adatállományában Massachusetts állam kormányzóját és jutott hozzá egészségügyi adataihoz. A születési dátum, az irányítószám és a nem alapján a regisztrált választópolgárok publikus adatbázisával sikeresen össze tudta kapcsolni az egészségügyi adatokat tartalmazó adatállományt. 2006-ban feltörték az AOL (American Online) által publikált anonim adatállományt, amely a webes keresőprogramba begépelt szövegeket tartalmazta. Ugyancsak 2006-ban feltörték a Netflix videó kölcsönző hálózat anonim adatállományát, amelyben a mozifilmek nézői értékelése szerepelt. Utóbbi adatállományt két matematikus, Narayanan and Shmatikov törte fel és cikket is írtak a munkájukról [2].

Az Egyesült Államokban nincs szövetségi adatvédelmi törvény, azonban 1996-ban szövetségi egészségügyi adatvédelmi törvényt hoztak létre (Health Insurance Portability and Accountability Act-HIPAA), amelynek a függelékében ismertettek egy anonimizálási módszert, amelynek alkalmazása esetén elhanyagolható újraazonosítási kockázattal kell számolni. Ennek a megalkotásához figyelembe vették L. Sweeny [3] és P. Golle [4] kutatásait az amerikai lakosság földrajzi és életkori eloszlásával kapcsolatban. A kutatásokhoz mindketten az USA publikus népszámlálási adatállomá-

nyát használták fel. Az USA felismerte azt, hogy az egészségügyi adatokon a nem kellő körültekintéssel végrehajtott anonimizálás nemzetbiztonsági kockázatot jelenthet. Az ún. privacy rule [5] szerinti anonimizálás úgy történik, hogy az adatállományból minden azonosító számot, jelet eltávolítanak; az érintettel kapcsolatos dátumokból (születés, halál, beutalás, felvétel, elbocsátás stb.) csak az évszámot hagyják meg. A születési dátum esetében, a 90 évnél idősebb érintetteknél az évszám helyett a „90 évnél idősebb" szerepelhet csupán. Az érintettre utaló azonosítók 18 kategóriáját sorolja fel a HIPAA törvény nem kizárólagos módon, pl. név; cím (a címből csak egy 20 ezernél nagyobb lakosságra mutató irányítószám prefix maradhat meg, az ötjegyű irányítószámból legfeljebb a három első számjegy, ha a lakosok száma kevesebb, mint 20 ezer, akkor a háromjegyű irányítószám prefixet helyettesíteni kell 000-val)., telefon, fax, e-mail, egészségbiztosítási azonosító, orvosi naplószám, igazolások, engedélyek száma, gépek azonosítói, gyári számok, biometrikus adatok, arcot is ábrázoló fényképek stb.

A privacy rule meglepően jól vizsgázott a különböző feltörési kísérletekben. Peter Kwok [6] megpróbálta egy piackutatási adatbázissal összekapcsolni egy a privacy rule alapján anonimizált egészségügyi adatállományt, az illetékes etikai bizottság engedélyével, és 15 ezer orvosi rekordból kettőt sikerült névvel, címmel beazonosítania, ami 2 / 15 000 = 0,013% kockázat. Benitez and Malin [7] ugyancsak tesztelte a privacy rule hatékonyságát. Több támadási esetet vizsgáltak meg és 0,01-0,25% közötti kockázatokat mértek.

AZ ÚJRA-AZONOSÍTÁSSAL KAPCSOLATOS KORÁBBI EREDMÉNYEK

Először Sweeney [3] dolgozta fel az USA 1990-es népszámlálási adatait abból a célból, hogy megvizsgálja a nem, születési dátum és az irányítószám adatok eloszlását, és megbecsülje, hogy ilyen adatok ismeretében az Amerikai Egyesült Államok lakossága milyen arányban azonosítható. A pontos születési dátum nem volt elérhető, de feltételezte, hogy születési dátumok egyenletesen oszlanak el egy éven belül. Számításai alapján (meglehetősen durva becsléseket használt, amelyeket később, pontosabb mérésekkel nem sikerült igazolni) az USA lakosainak 87.1%-át egyértelműen azonosítja e három adat.

Az eredményeket szerette volna ellenőrizni, és az újabb, 2000-es népszámlálási adatokkal összevetni Golle [4]. Számára sem volt elérhető a pontos születési dátum. Azonban kiszámította, hogy egyenletes eloszlást feltételezve, várhatóan hány személy született különböző napon, hány olyan pár volt, akik azonos napon születtek stb. Egy n személyből álló csoportból az egyéneket véletlenszerűen b számú alcsoportba (dobozba) helyezzük (pl. születések dátuma éven belül), akkor az i embert tartalmazó alcsoportok száma a következő:

$$f_n(i) = \binom{n}{i} b^{1-n} (b-1)^{n-i} \qquad (1)$$

Egy, az újraazonosítással kapcsolatos fogalom a g-különböző (g-Distinct):

Egy személyt egykének nevezünk, ha a tulajdonságok egy olyan halmazával rendelkezik, amilyen senki másnak sincs. Azt mondjuk, hogy egy személy g-különböző, ha a tulajdonságait tekintve megkülönböztethetetlen g-1 vagy kevesebb személytől. Az egyediség azonos az 1-különböző fogalommal. A g-különböző személyek száma $h_n(g)$ az *1, 2, …, g* személyt tartalmazó alcsoportokban (dobozban) található személyek számának összege.

Tegyük fel, hogy 70 személy ugyanabban az évben (nem szökőévben) született, akkor várhatóan $70 \times 365^{-69} \times 364^{69} \approx 57{,}93$ személynek különböző lesz a születésnapja, és $\frac{70 \times 69}{2} \times 365^{-69} \times 364^{68} \approx 5{,}49$ napon lesz egyszerre két személynek is születésnapja. Folytatva a számolást, tehát a 70 emberből 57,93 (82,76%) lesz 1-különböző, és (98,44%) személy lesz 2-különböző. Egy olyan irányítószám körzetben, ahol 8-10 ezren laknak, azaz nagyjából 4-5 ezer férfi és 4-5 ezer nő él, közöttük 5000/70 ≈ 70 lesz olyan, akiknek ugyanabban az évben van a születésnapja.

k-Iker (k-Twin):

Tetszőleges k egész számra, ha egy adattábla *pontosan k* olyan személyt (rekordot) tartalmaz, amelyeknek azonosak a kvázi-azonosítóik (pl. irányítószám, születési dátum, nem), akkor őket k-ikreknek nevezzük. Minden olyan adatot, amely később elvben személyazonosításra szolgálhat, kvázi-azonosítónak nevezünk (1., 2. táblázat).

Irányítószám	Lakosság	1-iker	2-iker	3-iker	4-iker	5-iker	6-iker
6500	32660	18705	4997	1072	155	25	
4060	17795	12945	2065	213	19	<10	
6237	8829	7473	613	38	<10		
6635	4699	4331	181	<10			
8248	2969	2792	84	<10			
8096	1306	1272	17				
7381	817	807	<10				

1. táblázat
Példák a k-ikrek számára a népességnyilvántartásból

Irányítószám	Lakosság	1-iker	2-iker	3-iker	4-iker	5-iker	6-iker
6500	32660	19039.98	5029.45	957.79	143.76	17.95	1.93
4060	17795	13240.45	1936.18	202.29	16.57	1.12	
6237	8829	7629.82	554.96	28.51	1.14	0.04	
6635	4699	4347.56	166.39	4.43	0.09		
8248	2969	2828.17	67.25	1.09			
8096	1306	1271.14	12.32	0.08			
7381	817	806.98	4.49				

2. táblázat
A k-ikrek várható száma P. Golle képletével (1) számítva

A szerző néhány irányítószám körzetre meghatározta a népesség-nyilvántartástól kapott adatbázis alapján a k-ikrek számát (k = 1, ..., 6) értékekre, lásd az 1. táblázatot. A 2. táblázatban a Golle képlettel (1) kiszámított értékeket mutatja be. A számok meglepően hasonlók a két táblázatban. Golle az (1) képlettel kiszámította az USA-ban élő 1-különböző személyek számát minden egyes irányítószám körzetben,

majd az egész USA-ra összesítve azt találta, hogy a lakosság 63%-a 1-különböző, azaz egyértelműen azonosítható az irányítószám, a nem és a pontos születési dátum segítségével.

A TÉTELES EGÉSZSÉGÜGYI ADATTÁR

A Tételes Egészségügyi Adattár (TEA) [8] forrása az Országos Egészségbiztosítási Pénztár (OEP) három elszámolási adatállománya volt (gyógyszertárak vényjelentései, járóbeteg-elszámolás, fekvőbeteg-elszámolás). Létrehozására az Egészségügyi, Szociális és Családügyi Miniszter 2004/76. (VIII. 28.) számú rendelete alapján került sor 2004-ben. Az adatállományokban TAJ-t egy pszeudonimmel helyettesítették adatvédelmi okok miatt. A TAJ egy 9-jegyű digitális azonosító, az azt helyettesítő ún. pszeudo-TAJ ugyancsak egy 9-jegyű szám. Az egészségügyi átszervezések folytán a fogadó szervezetek száma napjainkra négyről kettőre csökkent: a Gyógyszerészeti és Egészségügyi Minőség- és Szervezetfejlesztési Intézetre [9], valamint az Állami Népegészségügyi és Tisztiorvosi Szolgálatra [10].

Ismert adatvédelmi problémák a Tételes Egészségügyi Adattárral kapcsolatban

A Tételes Egészségügyi Adattárat a miniszteri rendelet anonimnak nyilvánítja, holott valójában nem az. Összevetve a HIPAA privacy rule szabályaival, az adattár tartalmazza az intézmények, szervezeti egységek kódjait, orvosi naplószámokat, a kezelő és a beutaló orvosok azonosítóit, pontos születési, halálozási, felvételi, elbocsátási és beutalási, gyógyszerkiváltási dátumokat, irányítószámokat, az alkalmazott egészségügyi ellátás kódjait. Ezek mind kvázi-azonosítók és személyazonosításra használhatók. Az országos orvos nyilvántartó honlapon [11] a pecsétkódot megadva megtalálható minden működési engedéllyel rendelkező orvos adata, szakorvosi képesítései, munkahelye. A TEA adatbázisban a pro famila ellátások adatai külön meg vannak jelölve, ezért az orvosok családtagjai közvetlenül és egyértelműen azonosíthatók.

A szerző a TEA ügyében 2006-ban az Alkotmánybírósághoz fordult, de indítványát elutasították. Az AB kijelentette, hogy az OEP továbbra sem továbbíthat személyazonosításra alkalmas adatokat, de maga úgy ítélte meg, hogy a TAJ nélkül továbbított adatok nem alkalmasak személyazonosításra, azaz semmilyen adatvédelmi probléma nem merül fel. A TEA adatbázis – mivel anonim – ezért nem áll etikai bizottság felügyelete alatt sem.

AZ ÚJRA-AZONOSÍTÁS KOCKÁZATA

A korábban említett kutatások nem támaszkodhattak pontos születési adatokat tartalmazó adatállományokra. Az azonos napon született állampolgárok várható eloszlását a valószínűség-számítás segítségével becsülték meg. A szerző ellenben hozzájutott a magyar népesség-nyilvántartás eloszlási adataihoz, amellyel pontosabb kockázatmérésre nyílt lehetőség.

A népesség-nyilvántartás kutatási adatállománya

Az adatállomány 270MB méretű szöveges állományt jelent, amelynek felépítése az 1. ábrán látható.

```
...
6188;1994.04.27.;N;1
6188;1994.04.29.;N;1
6188;1994.05.03.;N;1
6188;1994.05.29.;F;1
6188;1994.06.18.;F;1

...
```

1. ábra
A népesség-nyilvántartástól kapott kutatási adatbázis felépítése (Az eredeti értékeket a szerző megváltoztatta, nincs 6188 irányítószám Magyarországon).

Minden egyes sor négy adatelemet tartalmazott pontosvesszővel elválasztva: az irányítószámot, a születési dátumot, a nemet (N-nő, F-férfi), és az ezen a napon született állampolgárok számát a megadott irányítószám körzetben.

Az újra-azonosítás kockázatának kiszámítása

Az újra-azonosítás kockázatának kiszámítása az említett k-iker fogalmon alapul. Egy k-iker halmazban pontosan *k,* valamilyen szempontrendszer alapján megkülönböztethetetlen személy található. Például az egy irányítószám körzetben, egy napon született, azonos nemű személyek (ha nem áll rendelkezésre más kvázi-azonosító), akkor megkülönböztethetetlenek.

Az újra-azonosítás kockázata az alábbi képlettel számolható:

$$kockázat = \frac{azonosítható\ személyek\ száma}{összes\ személy} \quad (2)$$

A k-különböző személyekre kiszámított kockázat a következő:

$$kockázat(k\text{-különböző}) = \frac{azonosítható\ személyek}{összes\ személy} = \frac{\sum_{1 \le i \le k} i \times number\ of\ (i\text{-iker})}{összes\ személy} \quad (3)$$

A (3) számú „pesszimista” képlet feltételezi, hogy a k-különböző személyek is valamilyen külső tudás segítségével azonosíthatók. Például az adatokat feltörni igyekvő tudja a célszemélyről, hogy a Kútvölgyi Szanatóriumban kezelteti magát, vagy tudja egy orvosi vizsgálat vagy műtét dátumát, tudja egy kiváltott gyógyszer nevét és dátumát. Ezzel a tudással azonnal azonosíthatóvá válik a célszemély akkor is, ha a demográfiai adatok alapján két (vagy több) lehetséges személy található az adatbázisban. A TEA adattár tele van kvázi-azonosítókkal ezért ennek a képletnek van létjogosultsága. Egyébként azonban valószínűleg egy másik képletet kell használni, amely figyelembe veszi azt, hogy *k* megkülönböztethetetlen személy közül a célszemélyt csak 1/k valószínűséggel lehet azonosítani. Ezzel a „realisztikus” (3*)

képlettel a kockázatot a következőképpen lehetne kiszámítani:

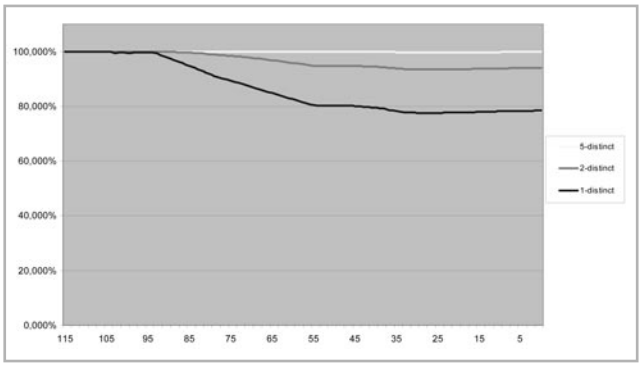
$$\text{kockázat}(k\text{-különböző}) = \frac{\sum_{1 \le i \le k} i \times number\ of\ (i\text{-iker}) \times \frac{1}{i}}{összes\ személy} = \frac{\sum_{1 \le i \le k} number\ of\ (i\text{-iker})}{összes\ személy} \quad (3^*)$$

A népesség-nyilvántartótól kapott adatállomány segítségével meghatározható a magyar lakosságban a k-ikrek száma, lásd 3. táblázat. A teljes lakosság 11-különböző (az USA lakossága 31-különböző volt). Az újra-azonosítás kockázata a „pesszimista" (3) képlettel számolva: 78.426%, 94.001%, and 99.801%; ha a (3*) képletet használjuk, akkor 78.426%, 86,214%, and 87,985% az 1-különböző, 2-különböző és az 5-különböző személyekre vonatkozóan.

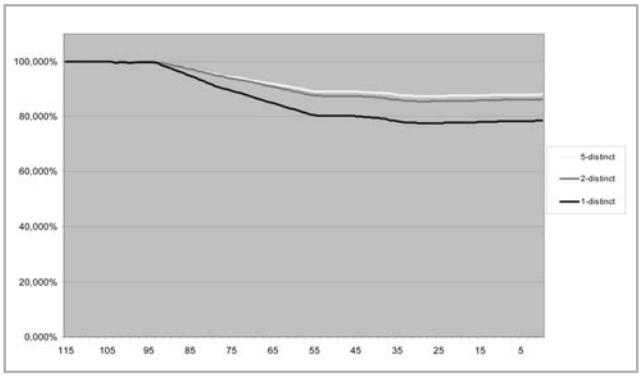
A k-ikrek száma a magyar lakosságban						
k	1	2	3	4	5	6
A k-ikrek száma	7 845 850	779 027	136 968	31 905	8 353	2 316
k	7	8	9	10	11	12
A k-ikrek száma	629	135	43	12	1	0

3. táblázat
A k-ikrek száma a magyar lakosságban, a 4-jegyű irányítószám, a születési dátum és a nem alapján

A teljes lakosságból 7 845 850 fő egyedi irányítószám, születési dátum, és nem adatokkal rendelkezik, de ha bármely két személy közül ki tudom választani a célszemélyt, akkor már 9 403 904 magyar állampolgár azonosítható az adatok alapján.



2. ábra
Az azonosítás kockázata a megadott életkornál idősebb populáció körében a "pesszimista" (3) képlettel számolva



3. ábra
Az újra-azonosítás kockázata az életkor függvényében a "realisztikus" (3*) képlettel számolva

Ha az életkor szerinti eloszlását vizsgáljuk meg a kockázatnak, akkor azt tapasztaljuk, hogy az idősebbek egyre nagyobb valószínűséggel azonosíthatók. A születések számának jelentős csökkenése miatt, a kockázat kisebb mértékben növekvő tendenciát mutat az 1980-as évektől kezdve napjainkig. Az újra-azonosítás kockázatát az életkor függvényében a 2. és a 3. ábra mutatja be. Három függvény mutatja az 1-különböző, a 2-különböző és az 5-különböző személyek azonosítási kockázatát a „pesszimista" és a „realisztikus" képlettel. Mindhárom függvény 100%-ról indul, és lassan csökken. A 76 évesnél idősebbek több mint 90%-a 1-különböző.

Az újra-azonosítás kockázatának csökkentése általánosítással

A népesség-nyilvántartás adatállománya lehetőséget adott arra, hogy megvizsgáljunk bizonyos általánosítási lehetőségeket, amelyek az újra-azonosítási kockázat csökkenhetik. Nemzetközi példák alapján a szerző megvizsgálta az: irányítószám első három jegye, az irányítószám első két jegye, a születési év és hónap, a születési év, az irányítószám első három jegye + a születési év és hónap, valamint az irányítószám első három jegye + születési év általánosításokat. Az eredményeket a 4. táblázat foglalja össze. Látható, hogy a pontos születési dátum alkalmazása általában jelentős kockázatot jelent, és hogy kizárólag a születés évét meghagyva a kockázat jelentősen csökkenthető. A legjobb eredmény két általánosítási transzformáció együttes alkalmazásával érhető el. A születési év + irányítószám első 3 jegye esetében a kockázat már 0.1% alá csökkent, ami még mindig 10 ezer egyértelműen azonosítható idős magyar állampolgárt jelent.

Általánosítás	1-különböző	5-különböző (pesszimista)	5-különböző (realisztikus)
Irányítószám első három számjegye	57.859%	98.711%	74.882%
Irányítószám első két számjegye	14.814%	71.286%	34.541%
Születési év és hónap	14.995%	50.962%	27.645%
Születési év	0.586%	6.244%	2.291%
Irányítószám első három számjegye + születési év, hónap	1.853%	18.183%	6.830%
Irányítószám első három számjegye + születési év	0.037%	0.274%	0.108%

4. táblázat
A különböző általánosítások segítségével kapott kockázatcsökkentés

ÖSSZEFOGLALÁS

A magyar egészségügyi kormányzat 2004-ben létrehozta a Tételes Egészségügyi Adattárat. Ez pszeudonimizált adatokat tárol a teljes lakosságról 1998-tól kezdve. Az OEP az elszámolási adatbázisokban szereplő TAJ azonosítókat rendre pszeudo-TAJ-ra cseréli ki, egyebekben azonban változatlan formában továbbítja azokat. Az ilyen módon pszeudonimizált adatállományban szerepel a páciensek születési

dátuma, neme és a lakóhelyük irányítószáma. E demográfiai adatok kezelésére korlátlan jogot kapott a GYEMSZI és az ÁNTSZ az Eüaktv. módosítása nyomán 2013-ban.

Az adatvédelmi biztos 2006-ben ellenezte a TEA adattár létrehozását, mert kockázatosnak ítélte a működést, kifejezetten veszélyesnek tartotta az állampolgárok háborítatlan magánéletére nézve [12]. Az adattár tartalma, a dátumok, az intézmények és orvosok adatai, a páciensek lakóhely és születési adatai kifejezetten alkalmasak arra, hogy egyes pácienseket azonosítsanak az adatok segítségével. Különösen az jelent problémát, ha a TEA adattár elhagyja a GYEMSZI és az ÁNTSZ területét, hiszen anonim adatokról van szó, és attól kezdve követhetetlen a felhasználók köre. A TEA adattáron végzett feldolgozási műveletek felett nem őrködik etikai bizottság, nincs független adatvédelmi felügyelet, és társadalmi kontroll sem az adatok hasznosításakor. Az adattár felhasználása nem a társadalom szeme előtt, hanem eltitkolva történik.

A szerző megállapítása szerint a magyar lakosság 78,4%-a a születési dátum, nem és irányítószám adatok alapján egyértelműen azonosítható. Ha valamilyen extra tudással is rendelkezünk, akkor az azonosítási kockázat azonnal 94% fölé emelkedik. Az nyugdíjas korú emberek, vagy a faluban, kisvárosban élők esetén ez a kockázat eleve 90% feletti. A miniszterelnök vagy az adatvédelmi hatóság elnöke például 1-iker, azaz egyértelműen azonosítható. A parlamenti képviselők a vezető politikusok, az Alkotmánybíróság tagjai kötelesek közzé tenni életrajzukat és vagyonnyilatkozatukat. E kettőből sokszor kideríthető a születésük dátuma és a lakóhelyük irányítószáma. Az ún. szociális hálózatokban (Facebook, Skype, iWiW, Yahoo, Google stb.) éppen ezeket az adatokat szokták megadni a felhasználók, amivel gyakorlatilag kulcsot adtak a TEA állomány feltöréséhez. Ezek gyakorlatilag elérhető adatok mindenkiről és ezért a TEA adatbázis működése komoly adatvédelmi aggályokat vet fel. Nem zárható ki az sem, hogy az adatbázis egyes részei külföldre kerültek, ami nemzetbiztonsági kockázatot jelenthet. Az állomány messze nem tekinthető tisztességesen anonimizáltnak.

IRODALOMJEGYZÉK

[1] Ohm, P., Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, (August 13, 2009). UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: http://ssrn.com/abstract=1450006

[2] Narayanan, A. and Shmatikov, V.: Robust De-Anonymization of Large Sparse Datasets, in Proceedings of the 2008 IEEE symposium on Security and Privacy, pp. 111-121, 2008

[3] Sweeney, L., Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000. URL: http://dataprivacylab.org/projects/identifiability/paper1.pdf utolsó letöltés: 2013. szeptember 15.

[4] Golle, P. Revisiting the uniqueness of simple demographics in the US population, in Proceedings of the 5th ACM workshop on Privacy in electronic society, pp. 77-80. ACM, 2006.

[5] Office for Civil Rights, Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule November 26, 2012, URL: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf, utolsó letöltés: 2013. szeptember 15.

[6] Kwok P, Davern M, Hair E, Lafky D (2011) Harder than you think: a case study of re-identification risk of HIPAA-compliant records. Chicago: NORC at The University of Chicago, Abstract #302255.

[7] Benitez, K., and Malin, B. Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association, Vol. 17 No. 2 (2010), pp. 169-177, doi:10.1136/jamia.2009.000026

[8] http://adatgyujtes.gyemszi.hu/TEA/

[9] http://www.gyemszi.hu

[10] http://www.antsz.hu

[11] A szolgáltatás az EEKH honlapjáról érhető el: http://kereso.eekh.hu/

[12] Az Adatvédelmi Biztos 1301/A/2006-9. számú állásfoglalása, 2006. október 9.

A SZERZŐ BEMUTATÁSA

Dr. Alexin Zoltán matematikusként végzett a József Attila Tudományegyetemen 1985-ben. Doktori fokozatát 2003-ban szerezte tanuló algoritmusok alkalmazásairól írt értekezésével. A SZOTE-PACS rendszer tervezése volt első orvosi informatikai feladata 1995-ben. 2004-ben kezdett el egészségügyi adatvédelemmel foglalkozni. Több alkotmánybírósági és más peres eljárást indított az egészségügyi adatkezelés jogi alapjainak tisztázása érdekében.

Szakértő volt az EuroSOCAP (European Standards on Confidentiality and Privacy in Healthcare) FP6 projektben. 2009-től a Dél-alföldi Regionális Kutatásetikai Bizottság tagja. 2009 és 2010 között közös adatvédelmi kutatásban vett részt a Central Lancashire egyetemmel. 2012-től a FutureICT.hu TÁMOP project adatvédelmi alprojektjének vezetője.