14.3.2018

Antti Honkela
Tilastotieteen apulaisprofessori
Matematiikan ja tilastotieteen osasto, matemaattis-luonnontieteellinen tiedekunta, sekä
Kansanterveystieteen osasto, lääketieteellinen tiedekunta
Helsingin yliopisto

Asiantuntijapyyntönne HE 159/2017 vp

Lausunto Hallituksen esityksestä eduskunnalle laiksi sosiaali- ja terveystietojen toissijaisesta käytöstä sekä eräiksi siihen liittyviksi laeiksi

Kiitän sosiaali- ja terveyslautakuntaa mahdollisuudesta esittää lausunto laista sosiaali- ja terveystietojen toissijaisesta käytöstä. Keskityn lausunnossani tietokantojen anonymisointiin ja anonymisoidun tiedon käyttöön. Kyseessä on aktiivinen tutkimusala, joka on kehittynyt ja laajentunut viime vuosina voimakkaasti. Olen tutkinut aihetta noin 4 vuotta ja julkaissut useita artikkeleja, joissa esitellään uusia menetelmiä lääketieteellisen tiedon käyttöön vahvan anonymiteetin suojassa.

EU:n yleisen tietosuoja-asetuksen johdanto-osan kohta 26 määrittelee ei-anonyymeiksi henkilötiedoiksi tiedot, jotka voitaisiin lisätietoja käyttämällä yhdistää luonnolliseen henkilöön. Tämän pohjalta on selvää, että anonymisointiin ei riitä suorien tunnistetietojen piilottaminen, vaan siinä on tarkasteltava kaikkien tietojen tunnistettavuutta kokonaisuutena.

Anonymisointimenetelmien teoreettisesta perustasta

Tieteellisessä kirjallisuudessa esiintyy muutamia erilaisia määritelmiä anonymisoinnille. Monet klassiset menetelmät perustuvat ajatukseen, että tietokannasta erotetaan kunkin henkilön tiedoista ne, joita pidetään sensitiivisinä. Tiedot anonymisoidaan karkeistamalla näitä niin, että kyllin monella on aina samat arvot sensitiivisissä tiedoissa (*k*-anonymity). Tulosta voidaan vahvistaa erilaisilla lisäehdoilla (esim. *l*-diversity, *t*-closeness). Näiden menetelmien heikkoutena sosiaali- ja terveystietojen kanssa on, että usein koko tietokanta on sensitiivistä eikä sitä siksi voida mielekkäästi anonymisoida.

Viime vuosina on noussut suosioon joukko *differentiaaliseen yksityisyyteen* perustuvia menetelmiä, joilla on puhtaan anonymisoinnin lisäksi mahdollista tehdä analyysejä, joiden tuloksen voidaan taata olevan loukkaamatta kenenkään yksityisyyttä. Differentiaalinen yksityisyys soveltuu tilanteeseen, jossa henkilötietoja sisältävää tietokantaa käytetään syötteenä laskentaoperaatiossa, jonka tulos ei

saa loukata tietokannassa esiintyvien henkilöiden yksityisyyttä. Menetelmää voi soveltaa erilaisiin laskentaoperaatioihin yksinkertaisten tilastollisten tunnuslukujen laskemisesta koko tietokannan anonymisoidun version tuottamiseen ja tekoälymallin oppimiseen tietokannasta. Differentiaalisen yksityisyyden takaamiseksi vaaditaan, että laskennan tulos ei saa muuttua kuin vähän, jos yhden henkilön tiedot tietokannassa vaihdetaan mielivaltaisesti. Koska samat takeet taataan kaikille, ei lopputulos voi riippua voimakkaasti kenenkään yksittäisen henkilön tiedoista. Sallitun muutoksen määrä on valittavissa oleva parametri, jolla säädetään taattua yksityisyyden tasoa. Tiukempi yksityisyys tekee yleensä tuloksista epätarkempia ja siten vähemmän hyödyllisiä. Ensimmäisiä differentiaaliseen yksityisyyteen perustuvia menetelmiä on hiljattain otettu tai ollaan ottamassa käyttöön niin julkishallinnossa (US Census 2020[1]) kuin Internet-yrityksissäkin (mm. Google, Apple, Mozilla Firefox).

Anonymisoinnin rajoitukset

Tietojen täydellinen anonymisointi on yleisesti mahdollista vain, jos lopputulos ei sisällä mitään informaatiota alkuperäisestä tietokannasta. Tällöin anonymisoitu tieto on kuitenkin täysin hyödytöntä. Toimiva anonymisointi onkin aina jonkinasteinen tiedon hyödyllisyyden ja anonymiteettitakeiden välinen kompromissi, joka vaatii harkintaa tietokannan sensitiivisyyden ja saavutettavissa olevan hyödyn välillä.

Koska jokainen anonymisoitu tietokanta vuotaa informaatiota alkuperäisestä tietokannasta, voi useita samoista tiedoista anonymisoituja tietokantoja yhdistelemällä olla mahdollista murtaa tietojen anonymisointi. Tämä riski voi toteutua, jos samoja tietoja anonymisoidaan moneen kertaan eri käyttäjille tai erilaisia tietokantoja tai muuttujia yhdistellen, ja joku saa haltuunsa nämä erilaiset anonymisoidut tietokannat. Differentiaalisen yksityisyyden teoriassa ilmiö näkyy siten, että jokainen uusi samasta alkuperäisestä tietokannasta johdettu anonymisoitu tulos heikentää anonymiteettitakeita. Tiukasti tulkiten tämä asettaisi ylärajan sille, kuinka monta kertaa samoja tietoja voidaan käyttää eri anonymisoinneissa, mikä heikentäisi tietojen laajempia käyttömahdollisuuksia olennaisesti. Tämän ongelman välttämiseksi on syytä harkita keinoja rajoittaa anonymisoitujen tietokantojen yhdistelyä esimerkiksi edellyttämällä erityistä lupaa niiden julkaisuun ja varmistamalla käyttölupaviranomaiselle oikeus käyttää tietopyyntöjen käsittelyssä kokonaisharkintaa tietojen anonymiteetin säilymisen varmistamiseksi. Mahdollisuus anonymisoitujen tietojen luvanvaraiseen julkaisuun olisi kuitenkin tärkeä avoimen tieteen ja datan avoimuuden edistämiseksi, mutta edellä kuvatun vuoksi tätä varten olisi syytä luoda erillinen lupaprosessi.

Tietyn tietokannan anonymisoinnin helppous ja tehokkuus riippuu vahvasti sen tietosisällöstä. Kustakin henkilöstä vain yksittäisiä tietoja sisältävän tietokannan tehokas anonymisointi on usein melko helppoa, kun taas hyvin monipuolista tietoa kuten kokonaisia sairaskertomuksia sisältävän tietokannan anonymisointi kokonaisena todennäköisesti hävittää kaiken hyödyllisen informaation. Vaikka koko tietokannan anonymisointi ei ole mahdollista, monipuolista tietoa sisältävää tietokantaa voi käyttää anonymiteetin turvaavasti differentiaalisen yksityisyyden suojassa kohdennetumpien analyysien tekemiseen, kuten tilastollisten testien tekoon tai regressiomallin sovittamiseen. Näiden tehokkaassa toteutuksessa tarvitaan usein monia kyselyjä tietokantaan, joiden avulla mallia parannetaan vähän kerrallaan. Lakiesitykseen kannattaisikin lisätä mahdollisuus tarjota anonymisoitujen tietokantojen luovutuksen lisäksi anonymiteetin turvaava ohjelmointirajapinta suojattavien tietojen käsittelyyn, koska tämä mahdollistaisi anonymisoitujen tietojen paljon paremman hyödynnettävyyden

[1] Aref N. Dajani *et al*. The modernization of statistical disclosure limitation at the U.S. Census Bureau, 2017. https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf

tiukasta anonymisoinnista huolimatta. Tämän toteuttamiseksi kullekin tietopyynnölle voitaisiin määritellä oma yksityisyysbudjetti, jonka puitteissa rajapintaa käyttäviä kyselyjä voisi tehdä.

Anonymisoinnin lupakäsittelystä

Kuten edellä todettiin, tietojen anonymisointiin liittyy aina jonkinasteinen kompromissi tiedon hyödynnettävyyden ja anonymiteetin välillä, mikä vaatii eettistä harkintaa. Tämä tulee erityisesti esiin differentiaalisen yksityisyyden kohdalla, jossa on valittava kullekin tapaukselle yksityisyyden tasoa säätelevä parametri tai yleisemmin tiettyä hanketta vastaavan yksityisyysbudjetin koko. Yhdysvalloissa U.S. Census Bureau on päättänyt erottaa tämän eettisen tarkastelun järjestelmän teknisestä toteuttajasta erilliselle Data Stewardship Executive Policy Committee -elimelle[2]. Mielestäni Suomessakin on syytä harkita eri viiteryhmiä ja erilaista teknistä, yhteiskunnallista ja eettistä asiantuntemusta yhdistävän asiantuntijaelimen perustamista tekemään periaatepäätökset eri tapauksissa vaadittavasta anonymiteetin tasosta ja samalla sallitun yksityisyysbudjetin suuruudesta.

Yhteenveto suosituksista

Yllä esitetyn perusteella suosittelen lakiesitykseen seuraavia muutoksia:

1. Anonymisoitujen aineistojen luovutuksen lisäksi tulee säätää mahdollisuudesta käsitellä ja analysoida aineistoa yksityisyyden ja anonymiteetin turvaavan ohjelmointirajapinnan yli suorittamalla tietokantaan useita kyselyjä, jotka kokonaisuutena säilyttävät riittävän yksityisyyden ja anonymiteetin.
2. Rajoitetaan anonymisoitujen aineistojen edelleenlevitystä sen varmistamiseksi, ettei niitä yhdistelemällä voida loukata anonymiteettisuojaa. Säädetään käyttölupien myöntäjille oikeus kokonaisharkintaa käyttäen evätä tietopyynnöt, joiden on syytä epäillä voivan johtaa anonymiteettisuojan loukkauksiin.
3. Luodaan avoimen tieteen ja tiedon edistämiseksi menettely, jolla anonymisoituja aineistoja voidaan erillisen luvan perusteella saattaa julkisiksi.
4. Perustetaan erillinen laaja-alainen asiantuntijaelin, joka tekee periaatepäätökset eri tietokantojen anonymisoinnissa vaaditusta yksityisyyden tasosta mm. pyydettyjen tietojen sensitiivisyyden ja sovelluksen potentiaalisen hyödyn huomioiden.
5. Aineistojen luovutukseen liittyvien palvelujen hinnoittelussa huomioidaan akateemisten ja kaupallisten toimijoiden erilaiset taloudelliset resurssit. Erityisesti on syytä varmistaa, ettei palveluiden hinnoittelu muodostu akateemisen tutkimuksen esteeksi ja siten rajoita sen vapautta.

[2] https://privacytools.seas.harvard.edu/why-census-bureau-adopted-differential-privacy-2020-census-population