PRIVASA

Uudet yksityisyyttä turvaavat menetelmät mahdollistamaan tietopohjaista tutkimusta, kehittämistä ja päätöksentekoa

1. Yhteenveto

Data on raaka-aine, joka ruokkii tietopohjaista päätöksentekoa sekä tutkimus-, kehittämis- ja innovaatiotoimintaa. Esimerkiksi ihmisten terveyttä, kuluttamista ja liikkumista koskevat aineistot ovat yhteiskunnallisesti merkityksellisiä. Tutkijoiden, tuotekehittäjien ja päättäjien kiinnostuksen kohteina ovat yhä useammin yksilölliset valinnat ja haasteet, jolloin tilastollisten koosteiden tarkkuus ei vastaa heidän käyttötarpeitaan. Datan hyödyntäminen ei kuitenkaan saa vaarantaa ihmisten perusoikeuksia, kuten oikeutta yksityisyyteen. Mitä tarkempaa yksilötason tietoa aineistot sisältävät, sitä suurempia ovat myös niihin liittyvät tietosuojariskit. Tästä syystä yksilötason aineistojen käyttöön ja jakamiseen kohdistuu erityistä varovaisuutta.

Nykytilanteessa datavarantojen heikko hyödynnettävyys uhkaa erityisesti yrityslähtöistä TKI-toimintaa. Yritysten toimintaedellytysten kehittäminen on yhteiskunnallisesti ja kansantaloudellisesti arvokas tavoite, jonka yhtenä osa-alueena pidämme dataan perustuvan tuotekehityksen ja yhteiskehittämisen tukemista. Koottujen tilastotietojen ja aidon, yksilötason aineiston ja välillä on kuilu, jonka ylittäminen vaatii aikaa, resursseja ja osaamista. Synteettisen datan avulla voimme kaventaa tätä kuilua ja siirtyä asteittain kohti joustavampia yhteistyön muotoja, jotka myötäilevät tutkimuksen ja tuotekehityksen eri vaiheita. Myös tieteellinen tutkimus ja opetus hyötyisivät ajantasaisista, aidonkaltaisista opetus- ja testiaineistoista, joiden käyttö olisi aiempaa sujuvampaa.

Jotta synteettisen datan käyttö voitaisiin omaksua osaksi TKI-toiminnan prosesseja, tulee meidän lisätä ymmärrystämme sen eri käyttötarkoituksista ja niiden asettamista reunaehdoista. Esimerkiksi synteettiselle datalle asetettavat laatu- ja tietosuojakriteerit voivat poiketa toisistaan suurestikin riippuen siitä, käytetäänkö aineistoa avoimeen järjestelmätestaukseen vai alustavaan tutkimushypoteesien muotoiluun tietoturvallisessa käyttöympäristössä. Lisäksi synteettisen datan tuottamiseen on tarjolla useita erilaisia menetelmiä, jotka eivät automaattisesti takaa tuotetun aineiston tietosuojaa. Synteettisen datan tarjoamat tietosuojatakuut onkin aina arvioitava tapauskohtaisesti, suunniteltu käyttökonteksti huomioiden.

On myös syytä korostaa, että synteettistä dataa ei tule käyttää tilanteissa, joissa aidolla datalla saavutettava tarkkuus on ensiarvoisen tärkeää, esimerkiksi kliinisessä päätöksenteossa. Useimmissa tapauksissa synteettinen data on välivaihe, joka mahdollistaa tutkimus- tai kehittämistyön ensiaskeleet ja auttaa näin tunnistamaan mahdolliset haasteet ajoissa.

Tietopohjainen tutkimus, kehittämistoiminta ja päätöksenteko tarvitsevat selkeitä pelisääntöjä aineistojen hyödyntämiseen. Kuten uusien teknologioiden käyttöönotossa yleisesti, myös synteettisen datan hyödyntäminen edellyttää tutkimusta, konseptointia, testausta sekä vallitsevien toimintatapojen uudelleenarviointia. Uusiin teknologioihin, kuten generatiiviseen tekoälyyn, ei tule suhtautua ainoastaan tietosuojariskien aiheuttajina, vaan niiden arvo on tunnistettava myös tietosuojan vahvistajina. Mahdollistava lainsäädäntö on nostettu yhdeksi keskeisimmistä kehityskohteista esimerkiksi äskettäin julkaistussa kansallisessa terveysalan visiossa[1] ja yhteistyössä alan katto-organisaatioiden kanssa tuotetussa Sotedigin työkalupakissa[2].

Kaikki tämä on mahdollista tutkimuslaitosten, yritysten ja julkisten rekisterinpitäjien yhteistyönä. Näistä lähtökohdista käynnistyi vuonna 2021 PRIVASA-hanke, jonka oppeihin ja havaintoihin nämä politiikkasuositukset perustuvat. Kiitämme lämpimästi kaikkia yhteistyökumppaneita kolmivuotisesta matkasta.

Avainsanat: TKI-yhteistyö, terveysdata, koneoppiminen, tietosuoja, toisiokäyttö, synteettisen datan tuottaminen

2. Politiikkasuositukset

PRIVASA-hankkeen politiikkasuositukset:

1. Riskiperusteiset tietosuojan viitearvot luomaan selkeyttä ja ennustettavuutta.
2. Mahdollistavaa lainsäädäntöä tukemaan yrityslähtöistä innovointia.
3. Joustavia toimintamalleja tukemaan yritysten ja tutkimuslaitosten yhteistyötä, myös kansainvälisesti.
4. Synteettinen data osaksi kansallisen TKI-toiminnan strategista kehittämistä.

[1] Suomen terveysalan kasvun ja kilpailukyvyn visio 2030, Sitra työpaperi (2023)
[2] Sotedigin työkalupakki - Kohti vaikuttavampaa sosiaali- ja terveydenhuoltoa (2023)

Suositus 1: Riskiperusteiset tietosuojan viitearvot luomaan selkeyttä ja ennustettavuutta

Tarve: Tarvitsemme konkreettisia esimerkkejä siitä, miten riittävä tietosuojan taso voidaan todentaa rekisterinpitäjille ja tietosuojaviranomaisille. Tietosuojalaeissa esiintyvä kahtiajako anonyymiin ja ei-anonyymiin tietoon ei vastaa todellisuutta. Tietosuoja voidaan nähdä jatkumona, jonka toisessa päässä ovat matalan tietosuojariskin sisältävät aineistot ja toisessa korkean riskin aineistot. Toisin kuin perinteisestä lain tulkinnasta voisi päätellä, tapa, jolla aineisto on tuotettu, ei yksinään määrittele saavuttavan tietosuojan vahvuutta. Tästä seuraa, että lakien tulkintaan ja soveltamiseen liittyy epävarmuutta, joka hidastaa uusien, mahdollisesti tehokkaampien menetelmien käyttöönottoa.

Ratkaisu: Tietosuojan tasoa kuvaavat kvantitatiiviset mittarit, kuten differentiaalinen yksityisyys, edistäisivät kestävien datastrategioiden kehittämistä. Riskiperusteinen lähestymistapa voi kannustaa organisaatioita kehittämään uusia ja innovatiivisia tapoja ratkoa tietosuojahaasteita, edistäen samalla teknologista ja liiketoiminnallista kehitystä. Esimerkiksi julkishallinto voi omalla esimerkillään johtaa uusien käytäntöjen omaksumista. Havainnolliset mittarit edesauttaisivat myös läpinäkyvää viestintää rekisteröidyille, lisäten ymmärrystä riskeistä ja tarjoten sitä kautta parempia vaikutusmahdollisuuksia omien henkilötietojensa käsittelyyn.

Suositus 2: Mahdollistavaa lainsäädäntöä tukemaan yrityslähtöistä innovointia

Tarve: Yksilöllistettyjen tuotteiden ja palvelujen aikakaudella suomalaiset yritykset tarvitsevat kehitystoimintaansa yksilötasoista dataa. Nykyinen tietosuojalainsäädäntö jakaa toimijoita useaan leiriin. Toisessa ääripäässä ovat riskinottajayritykset, joiden datastrategia perustuu lain minimivaatimusten täyttämiseen: erityisesti monikansallisten suuryhtiöt ovat saaneet toistuvasti huomautuksia laittomista ja epäeettisistä tietosuojakäytännöistään. Toisessa ääripäässä ovat vastuullisuuteen panostavat yritykset ja julkishallinto, jotka toimivat riskejä välttäen. Suomella ei ole varaa TKI-toiminnan seisahtumiseen ja siirtymiseen ulkomaille, vaan vastuullisille toimijoille on turvattava mahdollisuudet yksilötasoisen terveysdatan hyödyntämiseen tietosuojaa kunnioittavalla tavalla.

Ratkaisu: Tietosuojatakuut sisältävä synteettinen data voisi tarjota yrityksille yksilötason tietoa nykyisten aggregoidun tilastotiedon sijaan. Lainsäädännössä tulisi tunnistaa uusien teknologioiden luomat mahdollisuudet. Joissakin tilanteissa huolellisesti suojattu yksilötasoinen data, kuten differentiaalista yksityisyyttä toteuttava synteettinen data, voi tarjota jopa vahvemman tietosuojan kuin alkuperäisestä aineistosta johdetut, käsittelemättömät summamuuttujat.

Suositus 3: Joustavia toimintamalleja tukemaan yritysten ja tutkimuslaitosten yhteistyötä, myös kansainvälisesti

Tarve: Perinteisten lupaprosessien rinnalle tarvitaan kevennettyjä vaihtoehtoja, jotka mahdollistavat matalan kynnyksen kokeilut ennen suurempia investointeja. Datalähtöisten ratkaisujen aiempaa joustavampi testaus- ja pilotointi edistäisi merkittävästi etenkin pienten ja keskisuurten yritysten kykyä tehdä yhteistyötä tutkimuslaitosten kanssa. Myös kansainvälisen yhteistyön merkitys korostuu entisestään, kiihtyvän teknologiakehityksen myötä. Näiden tukeminen edellyttää paitsi tarvittavien aineistojen nopeaa ja kustannustehokasta saatavuutta, myös tietoturvallisia menetelmiä aineistojen yhteiskäyttöön ja jakamiseen.

Ratkaisu: Yritysten saavutettavissa olevat synteettiset malliaineistot toimisivat kansanterveyttä edistävän tuote- ja palvelukehityksen vauhdittajina. Findatan vastikään julkaisemat valmisaineistot ovat esimerkki oikeansuuntaisesta kehityksestä, ja esitämme, että vastaavia aineistokokonaisuuksia voisi tuottaa ja jakaa (luvanvaraisesti) myös synteettisinä versioina, jolloin ne olisivat entistä laajemmin hyödynnettävissä. Tämäntyyppistä kehitystä ovat viime vuosina edistäneet mm. Clinical Practice Research Datalink[3] (Iso-Britannia ja Pohjois-Irlanti) ja U.S. Census Bureau[4] (Yhdysvallat). Erilaisten malliaineistojen käyttöä ja tunnettuutta voitaisiin tukea edelleen kehittämällä niiden ympärille sujuvia palvelupolkuja ja mahdollisuuksien mukaan myös testialustatoimintaa. Synteettisen datan hyödyt niin rekisterinpitäjille kuin aineiston käsittelijöillekin voivat kertautua käytön aikana, esimerkiksi jos synteettisen datan käyttö mahdollistaa yksinkertaisemmat aineistonhallinnan prosessit.

Suositus 4: Synteettinen data osaksi kansallisen TKI-toiminnan strategista kehittämistä

Tarve: Suomen kannattaa kartoittaa synteettisen datan soveltuvuutta kansallisen TKI-toiminnan välineeksi. Synteettisen datan tarjoamat mahdollisuudet tulisi huomioida erityisesti osana kansallisten rekisterien ja tietovarantojen hyödyntämistä. Vastaavaa kehitystyötä on maailmalla tehty jo pidempään. Esimerkiksi Research Data Scotland (Iso-Britannia ja Pohjois-Irlanti) johtaa aktiivista strategiatyötä synteettisen datan parissa. South Australian Health (SA Health, Australia) on äskettäin käynnistänyt yhteistyön Gretel AI -nimisen yrityksen kanssa kartoittaakseen synteettisen datan mahdollisuuksia. Yhdistyneiden kansakuntien Euroopan talouskomissio (UNECE) on tammikuussa 2023 julkaissut viralliset ohjeet[5] synteettisen datan käytöstä tilastotietojen tarjoamisen välineenä. Niin ikään talouspuolella Financial Conduct Authority (FCA, Iso-Britannia ja Pohjois-Irlanti) perusti maaliskuussa 2023 synteettiseen dataan keskittyvän asiantuntijaryhmän.

---

[3] https://www.cprd.com/
[4] https://www.census.gov/about/what/synthetic-data.html
[5] https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide

Ratkaisu: Synteettiseen dataan liittyviä kysymyksiä tulee ensisijaisesti nostaa osaksi olemassa olevien asiantuntijaryhmien toimintaa, mutta erityisesti myöhemmissä vaiheissa myös oman, monitieteisen asiantuntijaryhmän kokoaminen voi olla perustelua. Ehdotetut toimet tukisivat sitä, että Suomi säilyttää asemansa kansainvälisesti tunnustettuna, datalähtöistä päätöksentekoa, tutkimusta ja kehitystä tukevana mallimaana. Asia on erityisen ajankohtainen nyt, kun eurooppalaiset data-avaruudet tekevät tuloaan. Hajautetun analytiikan ohella markkinapotentiaalia löytynee synteettisestä datasta, ja luvassa on huomattavia mahdollisuuksia suomalaisille yrityksille, tutkimuslaitoksille ja julkishallinnon organisaatioille.

Synteettinen data on mallinnettua aineistoa, jota voidaan käyttää tietosuojan parantamiseen.

Tässä politiikkasuosituksessa tarkoitamme synteettisellä datalla mallinnettua aineistoa, joka jäljittelee todellisia havaintoja. Synteettiset aineistot noudattelevat todellisten havaintojen rakennetta ja riippuvuussuhteita. Usein synteettiset aineistot ovat alkuperäisen aineiston pohjalta mallinnettuja havaintojoukkoja.

Tietosuojan parantaminen on vain yksi monista syistä, joiden vuoksi synteettistä dataa tuotetaan. Kaikki synteettiset aineistot eivät siis sovellu jaettaviksi, vaan ne voivat sisältää alkuperäisessä aineistossa esiintyneitä arkaluontoisia tietoja. Synteettisten aineistojen tuottaminen tavalla, jolla turvataan alkuperäisessä aineistossa esiintyneiden henkilöiden tietosuoja, edellyttää tähän tarkoitukseen soveltuvia menetelmiä.

Tietosuojanäkökulmasta erityyppiset aineistot muodostavat jatkumon, jonka toisessa ääripäässä ovat alkuperäiset, henkilötietoja sisältävät aineistot. Pseydonymisoidusta aineistosta puuttuvat ilmeisimmät henkilöllisyyden paljastavat tiedot, kuten nimi ja sosiaaliturvatunnus. Anonymisointi pyrkii estämään myös henkilöiden epäsuora tunnistamisen tietoja yhdistelemällä, esimerkiksi syntymäkuukauden ja postinumeroalueen perusteella. Perinteisten anonymisointimenetelmien tarjoama tietosuoja on kuitenkin useissa yhteyksissä todettu riittämättömäksi. Jatkumon toista ääripäätä edustavat tekaistut aineistot, jotka eivät perustu millään tavalla todellisiin havaintoihin.

Mitä on differentiaalinen tietosuoja?

Differentiaalinen yksityisyys on matemaattinen kehys, joka parantaa tietosuojaa, kun dataa analysoidaan tai jaetaan. Differentiaalisen yksityisyyden tarjoama yksityisyydensuoja perustuu tilastolliseen kohinaan, jonka ansiosta yksittäisen henkilön tiedot eivät erotu analyysituloksia vertailtaessa. Analyysiin tai synteettisen datan tuottamiseen sovellettavaa tietosuojan tasoa voidaan säätää yksityisyysparametrilla. Suurempi kohina tarkoittaa parempaa yksityisyydensuojaa, mutta heikentää vastaavasti aineiston laatua, rajoittaen mahdollisia käyttökohteita. Toisin sanoen sama epätarkkuus, joka suojelee aineistossa esiintyvien henkilöiden yksityisyyttä, kasvattaa myös tilastollisen testauksen virhemarginaalia. Differentiaalisen yksityisyyden merkittävänä etuna ovat matemaattiset takuut, joiden avulla yksityisyydensuojan vahvuus voidaan kuvata kvantitatiivisesti. Vaikka tekniset yksityiskohdat eivät poista tilannekohtaisen harkinnan tarvetta tietosuojaan liittyvissä kysymyksissä, suuntaa-antavien raja-arvojen määrittely voisi tukea terveysdatan vastuullista hyödyntämistä.

3. Yhteystiedot

PRIVASA-hankkeen vastuulliset johtajat:

Antti Airola, apulaisprofessori, Tietotekniikan laitos, Turun yliopisto
Tapio Pahikkala, professori, Tietotekniikan laitos, Turun yliopisto

Turun Ammattikorkeakoulun osatoteutusta johtivat:

Elina Kontio, yliopettaja ja tutkimusvastaava, Terveysteknologia, Turun AMK
Mojtaba Jafaritadi, yliopettaja ja tutkimusvastaava, Terveysteknologia, Turun AMK

VTT:n osatoteutusta johtivat:

Mika Hilvo, Research Team Leader (Health Data Analytics), VTT
Hankkeen alussa VTT:n edustajat olivat Mark Van Gils ja Miguel Bordallo.