Katoavatko katoanalyysit?

Sosiaalilääketieteen yhdistyksen, Suomen Epidemiologisen seuran, Suomen Tilastoseuran, Suomen Väestötieteen yhdistyksen ja Westermarck-seuran järjestämä iltapäiväseminaari Helsingissä 7.5.2018.

Euroopan tietosuoja-asetusta (GDPR General Data Protection Regulation) valmisteltiin vuosikausia, ja sillä oli kahden vuoden siirtymäkausi. Siitä huolimatta määräaika yllätti monet. 25.5.2018 alkaen asetusta on noudatettava kaikissa jäsenmaissa. Osa maista on suhtautunut muutokseen myönteisesti, erityisesti silloin kun uusi tietosuoja-asetus ei muuta merkittävästi tämänhetkisiä tutkimuskäytäntöjä. Toisaalta joissakin jäsenmaissa on maalattu kauhukuvia terveys- ja sosiaalirekisterien keruun loppumisesta ja suostumuksen vaatimisesta kullekin tutkimukselle erikseen.

Professori Jani Erola Turun yliopistosta kuvasi sosiologian kehityskulkua aineiston näkökulmasta kolmena ajallisena kehitysvaiheena. Ensimmäisessä vaiheessa 1800-luvulla huomattiin, että sosiaalisia ilmiöitä tulee tarkastella yksilöä laajemmalla tasolla. Tässä vaiheessa väestöä koskevaa tietoa kerättiin Erolan termein ”anekdoottisesti” – toisin sanoen tutkimus perustui ensisijaisesti siihen mennessä kerättyihin ja muiden koostamiin aineistoihin. Aineisto oli tutkijan itse luotettavaksi arvioimaa parasta mahdollista siihen mennessä kerääntynyttä dataa, jota aineiston keräämisen sijaan pikemminkin tarkasteltiin. Tiedonkeruu oli epätarkkaa, koska aineistot olivat paitsi riittämättömiä, niin myös alun perin kerätty aivan toista tarkoitusta varten. Ne eivät siis vastanneet haluttuihin kysymyksiin ja antoivat pahimmillaan virheellisiä vastauksia.

Kehityskulun toisessa vaiheessa 1940–50-luvuilla alettiin kerätä survey-aineistoja. Tärkeä muutos aiempaan oli se, että tutkijat itse vastasivat kysymyksenasetteluista ja tiedonkeruusta. Tutkimukselle oli siis tarkoitus, jota varten kerättiin spesifistä dataa. Tiedonkeruu paranikin, vaikeuttavat kato ja vastausvirheet tulosten tulkintaa. Vaikka survey-menetelmissä voidaankin hyödyntää tilastollisia korjausmenetelmiä, eivät nämä toimi, mikäli jokin merkittävä yhteiskunnallinen ryhmä puuttuu kyselystä. Esimerkkinä tästä Erola mainitsi Helsingin Sanomien luokkakonekyselyn, jossa saatiin nopeasti 40 000 vastausta, mutta vähäisten vastauksien vuoksi alle 40-vuotiaiden miestyöntekijöiden tietoja ei voitu edes imputoida painokertoimin.

Kehityksen kolmas ja viimeisin vaihe on pienimmän häiriön tiedonkeruu. Tiedonkeruumenetelmässä käytetään aiemmin kerättyä dataa niin pitkälle kuin mahdollista ja sitä täydentämään kerätään lisätietoja vain välttämätön määrä. Perustiedot tulevat siis eri rekistereitä, ja muulla tavalla kerätään vain ne tiedot, joita ei (ainakaan vielä) rekisteröidä. Dataahan kerätään kaikkialla ja koko ajan. Tutkijan näkökulmasta tämä helpottaa kato- ja muistamisongelmia, mutta toisaalta tämä ”pienimmän häiriön tiedonkeruu” edellyttää tietosuojan entistä tarkkaavaisempaa varjelua. Tähän tarkoitukseen Euroopan tietosuoja-asetus pyrkii.

Rekisteritiedot ovat pysyviä, mutta mitä tarkoittaa tietosuoja-asetuksen vahva ja tarkka suostumusvaatimus itse kerätyn aineiston kohdalla? Tietosuojavaltuutetun toimiston mukaan vähimmäisvaatimus on kymmenen kysymyksen listaus, josta tulee käydä ilmi muun muassa tiedot siitä, mitä ja mihin aineistoa kerätään, kuka aineistoa kerää ja käsittelee, milloin aineisto hävitetään, kerätäänkö tietoa muualta ja millä oikeusperusteella keruu tapahtuu. Tutkijalle asetus saattaa tuoda lisähaasteita. Esimerkiksi tutkimuksen kestoa voi olla hankalaa määritellä. Entä miten tulkitaan oikeutta tulla unohdetuksi? Voidaanko jo kerättyjä tietoja käyttää uudestaan? Laajan suostumuksen vaatimus ei ainakaan lisää jo nyt matalaa vastausaktiivisuutta, joten riskinä on uusien katoryhmien ilmaantuminen. Erityisenä vaarana on, että asetusta tulkitaan liian tiukasti varmuuden vuoksi. Lisäksi maiden tulkinnat asetuksesta saattavat vaihdella. Kansallisesti ainakin Ruotsin ja Tanskan tulkinta on ollut libe-

raali, mutta tulkinnat saattavat olla erilaisia myös saman maan eri laitosten ja viranomaisten välillä. Sama koskee myös Suomea. Saattaa tosin olla, että tietosuoja-asetuksen pelko on ollut vain myrsky vesilasissa.

Annamari Lundqvist kertoi, miten THL:ssä on pyritty lisäämään vastausaktiivisuutta, koska esimerkiksi terveystarkastustutkimuksissa vastausaktiivisuus on laskenut jo kymmenillä prosenttiyksiköillä. Lundqvistin mukaan vastauksia saadaan enemmän silloin kun vastaaja kokee kysymykset itselleen tähdellisiksi. Kyselyn pitäisi olla lyhyt ja ytimekäs, helpot kysymykset olisi oltava heti kyselyn alussa, osoitteen pitäisi olla käsin kirjoitettu, allekirjoituksen näkyä saatekirjeessä ja lisäksi paperinen kysely kannattaisi painaa yksipuoleisena. Arkaluonteiset ja sukulaisia koskevat kysymykset vähentävät vastaamista. Sen sijaan vastaajan nimi tai allekirjoitus kyselyssä, paperin laatu tai koko, kyselyn värit tai kirjasintyylit eivät näyttäisi vaikuttavan vastaamiseen. Saatekirjeessä tärkeintä on mainita kyselyn luottamuksellisuus, mutta muu teksti on vastausaktiivisuuden kannalta tarpeetonta. Myös ennakkoilmoitus kyselystä ja karhu(t) jälkikäteen parantavat vastausintoa, samoin kuin rahallinen korvaus. Lundqvist toi myös esille sen, että eri ikäryhmiin kannattaisi kohdistaa erilaisia tavoittelukeinoja. Näistä tarvitaan tosin vielä lisää tutkimustietoa.

Olli Pietiläinen Helsingin yliopistosta jatkoikin siitä, miten juuri nuorten vastausaktiivisuutta saataisiin parannettua. Helsingin kaupungin työtekijöillä toteutetuissa Helsinki Health Study -tutkimuksissa vastausprosentti on keski-ikäisillä 82 prosenttia, mutta nuorilla vain 52 prosenttia. Vuoden 2017 tiedonkeruussa alle 40-vuotiaista otettiin lisäotos laajemman aineiston saamiseksi. Sukupuolten, toimialojen ja työsuhteen vakinaisuuden tai pituuden suhteen ei vastausaktiivisuudessa ollut eroa. Vähiten vastasivat alle 25-vuotiaat ja työntekijät, jotka olivat pienimmässä tulonneljänneksessä ja/tai osa-aikaisessa työsuhteessa, kun taas vastaavasti ylemmät tai keskitason toimihenkilöt ja ylimmässä tuloluokassa olevat vastasivat eniten. Myös kyselyn aikana tai sitä ennen vähintään kahden viikon sairaslomalla olleilla vastaaminen oli muita ryhmiä vähäisempää. Parhaimmat tulokset saadaan yhdistelemällä useita tiedonkeruutapoja esimerkkinä verkkokyselyt, lomakekyselyt, muistutukset ja lisäksi vielä puhelinkysely tavoittamattomiksi jääneille.

Oona Pentala-Nikulainen THL:stä kertoi kommenttipuheenvuorossaan koko Suomen näkökulmasta, miten vastaajilta saadaan suostumus rekisteritietojen yhdistämiseen. Kansallisen terveys-, hyvinvointi- ja palvelututkimus FinSoten vuoden 2017–2018 tiedonkeruu kattoi kaikki maakunnat, ja sen otos oli noin 60 000 yli 20-vuotiasta vastaajaa. Vastaamisaktiivisuutta vähensivät muun muassa THL:n tietosuojavuoto, sote-uudistuksen viivästymisestä johtuva sote-väsymys ja lisäksi kapitaatiolaskelmista aiheutunut kohu. Aktiivisen suostumuksen pyytäminen rekisteritietojen yhdistämiseen laski vastausprosenttia, sillä aiemmin rekisteritietojen yhdistämistä varten tarvittava suostumus pyydettiin passiivisesti. Vastausprosentti laski 50 prosentista 46 prosenttiin. Vain 57 prosenttia antoi suostumuksensa ja sen antamiseen suhtauduttiin epäilevämmin etenkin vanhemmissa ikäryhmissä. Nettivastauksissa suostumusprosentti oli korkea ja nuorimmilla vastaajilla jopa 80 prosenttia. Suostumus saatiin herkemmin korkeasti koulutetulta, terveemmiltä, terveyssuosituksia noudattavilta ja niiltä, joilla ei ollut toimintakyvyn rajoituksia.

Tommi Härkänen THL:stä jatkoi kertomalla esimerkkinä FINRISKI -väestöanalyysin kadosta 1980-luvulta aina tälle vuosikymmenelle ja katoon vaikuttamisen keinoista moni-imputointimenetelmällä. Kun vuonna 1982 miesten vastausaktiivisuus oli 79 prosenttia ja naisilla 85 prosenttia, osallistui terveystarkastustutkimukseen vuonna 2012 enää 55 prosenttia 25–64-vuotiaista miehistä ja 64 prosenttia samanikäisistä naisista. Härkänen viittasi puheenvuorossaan aiempien puhujien huomioihin siitä, että vastausaktiivisuuteen vaikuttaa muiden aineistojen tapaan vastaajan ikä, sukupuoli ja koulutustaso. Lisäksi hän totesi, että tulosten korjaaminen on välttämätöntä, jotta harhoilta vältyttäisiin. Keskeisiä menetelmiä ovat aiemmin mainitut painotusmenetelmä ja imputointi. Härkänen toteaa Terveys 2011 -tutkimuksen esimerkkien pohjalta, että erityisesti moni-imputoinnilla voidaan päästä hieman parempaan tulokseen.

Jukka Jokinen THL:stä täydensi aiempia esityksiä lyhyessä puheenvuorossaan koko maan

kattavista kohorteista väestötutkimuksessa eli valtakunnallisten rekisteriaineistojen käyttämisestä kyselyiden rinnalla. Jokisen mukaan Suomessa on paljon kattavasti kerättyjä rekisteritietoja (esimerkiksi perusterveydenhuollon käyntisyyt, KELA:n etuusrekisterit, tartuntatautirekisteri ja hoitoilmoitusrekisterit), mutta tietojen vaikuttavuudessa voisi olla hiomista. Ideaalitilanteessa tietoa voisi kerätä jo esimerkiksi ennen varsinaista sairastumista, jotta saataisiin tietoja koko ajalta oireettomasta taudinaiheuttajasta, sairastumiseen ja kuolemaan asti. Jokinen totesi esityksessään myös, että yksilö voisi suostumuksellaan verrata tietojaan muuhun väestöön, mikä herätti yleisössä keskustelua.

Vastauspuheenvuorossa Mika Gissler THL:stä korosti, että tutkijoiden kannalta tärkeintä on saada mahdollisimman täydelliset ja kattavat tiedot tutkimuskäyttöön. Liian usein tutkija kohtaa tietojen panttaamista tai tietojen laadun heikentämistä ”tietosuojasyistä”. Esimerkiksi Eurostatin terveyskyselytutkimuksessa sokeat ja silmälaseja käyttämättömät yhdistetään yhdeksi ryhmäksi, jotta sokeiden tietosuoja turvattaisiin. Samalla jää kuitenkin uupumaan tieto heidän toimintakyvyn rajoitteestaan. Tuhannen euron kysymys onkin, miten turvata tieteen vapaus ja yksilön oikeudet. Anonymisointi, pseudonymisointi ja aineistojen analysoiminen etäyhteyksin parantavat tietosuojaa ja yksityisyyden suojaa. Toinen avainsana on koulutus alkaen kandidaattiopiskelijoista ja päätyen jo varttuneempiin tieteilijöihin.

Ulla Ahlblad-Bordi THL:ssä kertoi tietosuoja-asetuksesta ja sen aiheuttamista muutoksista. Aiemmasta poiketen asetus on ensisijainen laki, jota vain täydennetään – vielä käsittelyssä olevalla – kansallisella tietosuojalailla. Pääperiaatteet eivät ole muuttuneet:

– Tiedonkeruusta ja säilyttämisestä on informoitava.
– Tietoja saa käyttää ainoastaan siihen tarkoitukseen, jonka takia ne on kerätty. Tosin tilastointi ja tutkimus ovat kuitenkin poikkeuksia ja tieteellisen tutkimuksen käsitettä on laajennettu. Jatkossa tieteellinen tieto voi siis eurooppalaisittain merkitä myös esimerkiksi kaupalliseen tarkoitukseen tuotettua materiaalia.
– Tietoja ei saa kerätä turhaan.
– Tietoja saa säilyttää vain niin kauan, kun niitä tarvitaan.
– Tietojen on oltava eheitä ja luottamuksellisia ja täsmällisiä.

Aiempaa tarkemmin on osoitettava, että asetuksen periaatteita on noudatettu eri käytännesäännöin, sertifikaatioin ja tietotilinpäätöksin. Uusista ja jatkuvista tietokeruista on laadittava vaikutusten arviointi mm. tietosuojavuotojen ja -loukkauksien estämiseksi.

Arkaluonteisten tietojen kerääminen on mahdollista vain ”nimenomaisella suostumuksella” joka voi olla esimerkiksi kirjallisesti tai sähköisesti allekirjoitettu tai tekstiviestivahvistuksella varmistettu. Vielä jää epäselväksi, mitä tämä nimenomainen suostumus tarkoittaa. Rekisteritutkimukset ovat tulkinnan mukaan myös mahdollisia, kunhan kansallinen lainsäädäntö sen sallii. Arkaluonteisia tietoja voi kuitenkin aina kerätä suostumuksella, jonka on oltava vapaaehtoinen, yksilöity, tietoinen ja yksiselitteinen. Tekemättä jättäminen tai vaikeneminen ei tarkoita suostumusta. Suostumuksen on oltava myös aktiivinen: esimerkiksi valmiiksi valittu ruutu ei voi olla suostumus. Kohtuuttomia ehtoja ei voi asettaa, eikä osallistumattomuudesta saa aiheutua kielteisiä seurauksia. Suostumuksessa olisi myös ennakoivasti kerrottava mahdolliset aineiston siirrot EU- ja ETA-maiden ulkopuolelle. Lisäksi suostumuksen peruuttamisen pitää olla yhtä helppoa kuin sen antaminen. Epäselväksi jää, voidaanko jatkossa hyödyntää vastaamattomien tietoja aineiston analyysin. Esityksen lopuksi Ulla Ahlblad-Bordi totesi, ettei katoanalyysien sallimisesta ole vielä yksimielisyyttä.

THL:n tietoturvapäällikkö Christian Jämsén kommentoi tietoturvallisuutta rikollisten liiketoimintana esitellen karmivia lukuja. Verkkorikollisuuden arvo 600 miljardia dollaria ja sen arvioidaan kymmenenkertaistuvan ensi vuosikymmenen puolenväliin mennessä. Varastetuilla tiedoilla voidaan tehdä rahaa ja tietomurtoyrityksiä kohdistuu palvelimille päivittäin. Aineistoihin kohdistuvia riskejä voivat olla inhimilliset virheet, mutta myös tietoisiin väärinkäyttöihin ja tietomurtoihin on varauduttava. Koska monen tietoteknisen laitteen ja järjestelmän lähtökohtana on tietoturvallisuuden sijasta käyttömukavuus, tietojen rikastaminen ja henkilöhistorioiden muodostaminen on teknisesti helppoa tai

jopa erittäin helppoa. Tieteelliseltä tutkimukselta vaaditaan samalla tavalla tuloksellisuutta, mutta sama koskee myös tietosuojaa. Tämä aiheuttaa jatkossa paljon enemmän suunnitelmallisuutta ja dokumentointia, johon tutkijoiden on mukauduttava.

Antoisan iltapäivän päätti paneelikeskustelu. Keskustelussa nousi esille monen huoli rekisteritutkimuksen tulevaisuudesta. Voidaanko jatkossa enää kerätä tutkimusaineistoja, joissa epäsuora tunnistaminen on mahdollista? Voiko tutkija vahingossa tehdä tietosuojarikoksen? Onko aineiston täysin anonymisointi edes mahdollista? Yleinen näkemys oli, ettei GDPR-hysteriaa pitäisi lietsoa. Monet kysymykset ovat vielä auki, mutta tutkimusaineistojen keruulle ja analysoinnille on annettu monta poikkeusta tietosuoja-asetuksessa. Näitä pitäisi nyt voida hyödyntää siten, etteivät tutkijoiden mahdollisuudet kattavien ja laajojenkaan tietojen saantiin ainakaan kavennu – big datasta nyt puhumattakaan.

Laura Pääkkö
Opiskelija
Tampereen yliopisto

Mika Gissler
Tutkimusprofessori
Terveyden ja hyvinvoinnin laitos
Karoliininen instituutti