Ville Rissanen ja Erkka Nurmi

Polkuattribuutti-menetelmä harvinaisten prosessivarianttien anonymisointiin

Tietotekniikan pro gradu -tutkielma

3. Toukokuuta 2022

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Ville Rissanen ja Erkka Nurmi
Yhteystiedot: ville.rissanen@linux.com ja celetal@gmail.com
Ohjaajat: Timo Hämäläinen
Työn nimi: Polkuattribuutti-menetelmä harvinaisten prosessivarianttien anonymisointiin
Title in English: Path attribute -method for anonymizing uncommon process variants
Työ: Pro gradu -tutkielma
Opintosuunta: Koulutusteknologia ja ohjelmisto- ja tietoliikennetekniikka
Sivumäärä: 57+5
Tiivistelmä: Tutkielmassa selvitetään Suomen ja Euroopan Unionin lainsäädäntöä anonymisoinnin suhteen sekä yleisimpiä anonymisoinnin menetelmien toimintatapaa. Esitämme polkuattribuutti-menetelmän, jolla anonymisointi voidaan kohdentaa ainoastaan hoitopolun yksilöiviin osiin sen sijaan, että koko tapahtumarivi tai hoitopolku sensuroitaisiin osana anonymisointia. Menetelmän toimivuutta selvitetään kokeellisesti generoidulla datalla ja havaitaan, että polkuattribuutti-menetelmällä anonymisointu data korreloi hyvin vahvasti alkuperäisen aineiston kanssa.
Avainsanat: hoitopolku, polkuattribuutti, anonymisointi, pseudonymisointi, prosessi
Abstract: In the study we explore the legal requirements for anonymization in Finland and the European union. We also cover the mathematical basis and function of the most popular anonymization methods. We present a path attribute -method for anonymization where one can pinpoint anonymization to identifying nodes of the care pathway instead of censoring the entire event or the entire care pathway as a part of the anonymization process. We experiment with the validity of the method with generated data and we find that the data anonymized using the path attribute -method correlates strongly with the original data.
Keywords: care pathway, path attribute, anonymization, pseudonymization, process

Termiluettelo

Anonymisointi	Toimenpide tai joukko toimenpiteitä, joilla tieto muutetaan muotoon, josta yhdenkään yksittäisen ihmisen tunnistaminen ei ole realistisesti mahdollista ihmiskunnan tällä hetkellä tuntemalla teknologialla.
Pseudonymisointi	Anonymisoinnin tyyppi, jossa tilastosta sekä poistetaan assosiaatio koehenkilön ja hänen ominaisuuksiensa välillä että lisätään assosiaatio pseudonyymin ja ominaisuuksien välille. (ISO 25237:2017, kirjoittajan suomennos)
Pseudonyymi (subst.)	Luotu keinotekoinen tunniste, joka korvaa koehenkilön tunnistavat tiedot. Pseudonyymi ei linkity takaisin koehenkilön tunnistaviin tietoihin.
k-anonymiteetti	Aineiston ominaisuus, jonka toteutuessa ketä tahansa aineiston henkilöä ei voida tunnistaa k - 1 henkilöstä, joiden tiedot ovat aineistossa.
Tapahtuma	Yhtä todellisen maailman tapahtumaa kuvaava yhtenäisessä muodossa oleva lokitieto. Käytännössä tämän tulee sisältää ainakin tiedot siitä mitä on tapahtunut, koska ja kenelle.
Aktiviteetti	Asia, joka voi tapahtua. Prosessin vaihe ja tapahtuman “mitä”-tieto.
Prosessi	1. Joukko aktiviteetteja, joilla on tietty keskinäinen järjestys. Aktiviteettien välillä voi olla valintoja tai toistoja. 2. Toistuva tapahtumasarja
Prosessivariantti	Yksi mahdollinen vuo prosessin lävitse, sen alkupisteestä loppuun. Tutkielmassa variantit esitetään muodossa {[Ensimmäinen aktiviteetti]->[Toinen aktiviteetti]->...->[Viimeinen aktiviteetti]},

	jossa hakasulkujen sisältämät osat korvataan kunkin aktiviteetin tunnuksella.
Entiteetti	Olento, johon tietty joukko tapahtumia kohdistuu. Esimerkiksi auto, merikontti tai asiakas.
Prosessi-instanssi	Yksi toteutunut vuo prosessista. Esimerkiksi tietyn auton matka kokoonpanolinjalla tai tietyn asiakkaan palvelupolku. Yksi prosessi-instanssi toteuttaa aina yhden ja vain yhden prosessivariantin.
Palvelupolku	Joukko tapahtumia, jotka kohdistuvat tiettyyn entiteettiin. Saattaa sisältää useampia prosessi-instansseja ja myös useampia instansseja samasta prosessista. Esimerkiksi asiakkaan kaikkien asiointiprosessien luoma palvelupolku.
l-diversiteetti	*k*-anonymiteetin pidemmälle kehittävä aineiston anonymiteetin ominaisuus. Toteutuu kun kaikki salassapidon kannalta tärkeät arvot ovat vähintään hyvin edustettuna *l*:n verran arvoilla.
t-läheisyys	*l*-diversiteettiä pidemmälle kehittävä aineiston anonymiteetin ominaisuus. Toteutuu kun kaikki salassapidon kannalta tärkeät arvot ovat jakaumaltaan enintään t:n verran erossa toisistaan.
Differentiaalinen anonymiteetti	Matemaattinen kehys, jonka perusteella voidaan käyttää erinäisiä tilastollisen melun lisäämisen mekanismeja anonymisoinnin saavuttamiseksi siten, että yksittäisen henkilön tietoja on epätodennäköistä tunnistaa aineistosta.
Kohina	Tarkastelunäkökulman kannalta irrelevantti monipuolisuus aineistossa.

Kuviot

Taulukot

Sisältö

1 Johdanto

Terveysalan data-analytiikka edellyttää usein salassapidettävien terveystietojen hyödyntämistä tutkimuksessa, mutta terveysdata edelleen jakaminen aiheuttaa huolia eettisistä ja laillisista näkökulmista. Usein kiinnostavia eivät ole yksittäiset hoitokäynnit, vaan useammasta käynnistä koostuvat niin sanotut palvelupolut. Niitä tarkastelemalla saadaan esille koko hoitopolun erilaiset ilmentymät. Täten hoitopolkujen analysointi mahdollistaa koko prosessin kokonaisvaltaisen arvioinnin yksittäisten tapausten sijaaan. (Sweeney, 1997) (Kokkinakis & Thurin, 2007) (Elger, Iavindrasana, Lo Iocono, Müller, Roduit, Summers & Wright, 2009)

Euroopan unionin ja Suomen lainsäädäntö velvoittaa rekisterinpitäjiä pseudonymisoimaan ja/tai anonymisoimaan tiedon käsittelijälle päätyvän aineiston, mutta määritelmät siitä mikä on pseudonymisoinnin ja anonymisoinnin suhde keskenään sekä siitä mikä on yksityistietoa vaihtelevat eri instituutioiden ja tutkimusten välillä. Myös itse tietojärjestelmien ja tiedon jakamista toteuttavien asiantuntijoiden koulutus ja tietämys GDPR-lainsäädännöstä on havaittu puutteelliseksi. Lisäksi terveysalan yritysten tietosuojalausunnot eivät kaikki ole olleet GDPR-lainsäädännön kanssa yhteensopivia pitkään lainsäädännön voimaantulon jälkeen. (Alhazmi & Arachchilage, 2021) (Mulder, 2019)

Terveysalan tutkimuksellisesta näkökulmasta on tärkeää huomioida Suomen Tutkimuseettisen neuvottelukunnan (TENK) näkemys, jonka neuvottelukunta on laatinut yhteistyönä suomalaisen tiedeyhteisön kanssa. TENK:n tutkimuseettinen ohjeistus hyvistä tieteellisistä käytännöistä (HTK) ja niiden loukkausepäilyjen käsittelemisestä koskee luonnollisesti myös tutkimustoiminnassa hyödynnettävää data-analytiikkaa ja tutkimusaineiston käsittelyn käytänteitä ennen ja jälkeen tutkimusaineiston luomista tutkimustulosten data-analyysiä ja käsittelyä varten. (Tutkimuseettinen neuvottelukunta, 2013) (Tutkimuseettinen neuvottelukunta, 2019)

HTK-ohjeen tavoitteena on edistää kaikilta osin hyvää ja laadukasta tieteellistä käytäntöä (responsible conduct of research) ja samalla varmistaa, että myös mahdolliset tutkimusaineiston käsittelyyn liittyvät yksilön anonymiteettiä suojaavat toimenpiteet kyetään ehkäistä ennalta jo tutkimussuunnitelmia luotaessa. HTK-ohjeen mukaisesti tutkimus tulee suunnitella ja toteuttaa sekä raportoida siten, että syntyneet tietoaineistot tallennetaan tieteelliselle tiedolle asetettujen vaatimusten edellyttämällä tavalla. (Tutkimuseettinen neuvottelukunta, 2013) (Tutkimuseettinen neuvottelukunta, 2019)

Yksilön suojan näkökulmasta on korostuneesti sosiaali- ja terveysalan tutkimusaineistojen keräämis- ja koontivaiheessa on merkityksellistä, että tutkimuksissa sovelletaan tieteellisen tutkimuksen kriteerien mukaisia ja eettisesti kestäviä tiedonhankinta-, tutkimus- ja arviointimenetelmiä. Näin siitäkin huolimatta, että HTK-ohjeistuksen mukaisesti kaikissa tutkimuksissa pyritään noudattamaan tieteellisen tiedon luonteeseen kuuluvaa avoimuutta ja vastuullista tiedeviestintää tutkimuksen tuloksia julkaistaessa. Tutkimus suunnitellaan ja toteutetaan ja siitä raportoidaan sekä siinä syntyneet tietoaineistot tallennetaan tieteelliselle tiedolle asetettujen vaatimusten edellyttämällä tavalla. Yksilön mahdollista identifiointia tutkimusaineistossa tai anonymiteetin loukkaamista koskee HTK-ohjeistuksessa erityisesti mahdollinen piittaamattomuus tutkimustulosten tai tutkimusaineistojen puutteellinen kirjaaminen ja säilyttäminen. (Tutkimuseettinen neuvottelukunta, 2013) (Tutkimuseettinen neuvottelukunta, 2019)

Vaikka kunkin yksittäisen tapahtuman tunnistavat tiedot saataisiin anonymisoitua, voi olla mahdollista tunnistaa henkilö palvelupolun perusteella, jos samankaltaisia polkuja ei ole useita. (Kokkinakis & Thurin, 2007) (Elger et al., 2009) Terveysdatassa yksilöivät tekijät eivät ole aina ilmiselviä, joka ilmenee Rocherin, Hendrickxin ja De Montjoyen sekä Ravindran ja Graman tutkimusista. (Rocher, Hendrickx, De Montoye, 2019) (Ravindra ja Grama, 2021)

Deanonymisointihyökkäysten riskistä huolimatta suurin osa väestöstä ajattelee positiivisesti terveysdatan anonymisoidusta jakamisesta tutkimuskäyttöön. Harvinaisista sairauksista kärsivät joilla on tilastollisesti korkeampi riski deanonymisointiin ovat myös lähtökohtaisesti datan jakamisen puolesta, kunhan heidän yksityisyyttään ylläpitävät

toimenpiteet ovat riittävät. (Kalkman, van Delden, Banerjee, tyl, Mostert & van Thiel, 2022) (Courbier, Diamond, & Bros-Facer, 2019) (Karampela, Ouhbi & Isomursu, 2019)

Tämän tutkimuksen tarkoituksena on selvittää onko palvelupolkujen anonymisointi mahdollista ja kuinka kuinka toimivaa tällainen anonymisointi olisi. Lisäksi arvioimme oman menetelmämme pätevyyttä eri instituutioiden ja organisaatioiden määritelmien perusteella.

1.1 Yhteistyö tutkielmassa

Tutkielman laadinnassa työ oli jaettu karkeasti teoreettiseen ja kokeelliseen. Teoereettisesta osuudesta ja siihen liittyvästä kirjallisuuskatsauksesta vastasi Ville Rissanen ja kokeellisesta osuudesta ja siihen liittyvästä kirjallisuudesta vastasi Erkka Nurmi. Tästä huolimatta kumpikin kirjoittaja otti kantaa toisen osuuteen ja teki pienissä määrin muokkauksia. Kumpikin kirjoittajista osallistui johdannon, pohdinnan ja yhteenvedon kirjoittamiseen, jotka kirjoitettiin parityönä keskustellen. Johdantoon, pseudonymiteettiin ja anonymiteettiin liittyvä kirjallisuuskatsaus on tehty parityönä keskustellen. Työmäärä pyrittiin tasaamaan tunneissa puoliksi kirjoittajien kesken siinä määrin missä mahdollista.

2 Anonymisointi ja pseudonymisointi

Tässä kappaleessa tarkastellaan erinäisiä määritelmiä pseudonymisoinnille ja anonymisoinnille. Anonymisoinnin määritelmä on tutkielman kannalta tärkeä, koska se määrittää milloin tutkielman käsittelemä menetelmä täyttää anonymisoinnin asettamat vaatimukset.

Laajemmin tarkastellen myös potilastietojen, sairaskertomusten ja kaiken henkilöidentifikaatioon liittyvän tutkimuksellisen tietojen käsittelyn näkökulmasta anonymiteettiä koskevaa pohdintaa ja ohjeistusta avaa Tampereen yliopiston Tietoarkiston kehittämispäällikkö Arja Kuula-Luumi Vastuullinen tiede -verkkosivulla seuraavasti: “Yksi keino anonymiteetin varmistamiseen on taustatietojen kategorisointi. Esimerkinä Alavieskassa asuvasta tutkittavasta voidaan kertoa, että hän asuu maaseutumaisessa kunnassa Pohjois-Pohjanmaalla. Ryhmitykseen kuuluu 17 eri kuntaa ja tunnistaminen yksin iän, ammatin, sukupuolen ja luokitellun asuinalueen perusteella on jokseenkin mahdotonta.“ (Kuula-Luumi, 2018)

2.1 Pseudonymisoinnin määritelmä

Pseudonymisointi on ISO:n (International Organisation for Standardization) teknisen spesifikaation 25237:2017 määritelmän mukaisesti: “[A]nonymisoinnin tyyppi, jossa assosiaatio koehenkilön tunnistavien tietojen ja hänen ominaisuuksiensa välillä poistetaan ja korvataan uudella assosiaatiolla salanimeen tai muihin keinotekoisiin tekijöihin.” ISO 25237:2017 luokittelee pseudonymisoinnin anonymisoinnin alatyypiksi. (ISO 25237:2017, tutkielman kirjoittajien suomennos)

Euroopan Unionin tietosuoja-asetus (General Data Protection Regulation eli GDPR) määrittelee pseudonymisoinnin seuraavasti: “[H]enkilötietojen käsittelemistä siten, että henkilötietoja ei voida enää yhdistää tiettyyn rekisteröityyn käyttämättä lisätietoja, edellyttäen että tällaiset lisätiedot säilytetään erillään ja niihin sovelletaan teknisiä ja organisatorisia toimenpiteitä, joilla varmistetaan, ettei henkilötietojen yhdistämistä tunnistettuun tai tunnistettavissa olevaan luonnolliseen henkilöön tapahdu”. EU:n

määritelmän mukaan pseudonymisointi ei ole anonymisoinnin alatyyppi, jota käsitellään seuraavassa luvussa.

Itä-Suomen yliopiston tietosuojavastaava Helena Eronen määrittelee pseudonymisoinnin seuraavasti: “Pseudonymisointi tarkoittaa henkilötietojen käsittelemistä siten, että henkilötietoja ei voida enää yhdistää tiettyyn henkilöön ilman lisätietoja. Lisätiedot, esimerkiksi koodiavain, jonka avulla tunnistettavuus palautetaan, säilytetään erillään henkilötiedoista. Pseudonyymit tiedot ovat edelleen henkilötietoja ja niiden käsittelyyn sovelletaan tietosuojalainsäädäntöä.” (Eronen, 2019)

Täten huomaamme, että pseudonymisoinnin määritelmä vaihtelee eri organisaatioiden ja niiden tulkitsijoiden välillä. Myös pseudonymisoinnin suhde anonymisointiin on vähintäänkin kyseenalainen kahden merkittävän organisaation, ISO:n ja EU:n, ollessa eri mieltä näiden kahden suhteesta toisiinsa. Toinen tästä johtuva kiistanalainen tulkittava asianhaara on tuleeko pseudonymisoitu data luokitella yhä yksityistiedoksi. Tästä syystä tässä tutkielmassa otetaan huomioon lähinnä anonymisoinnin menetelmät, joiden toiminta anonymisoinnin työkaluna on yleisesti hyväksytty. Mutta itse anonymisoinnin määritelmä on myös kiistanalainen.

2.2 Anonymisoinnin määritelmä

ISO ja International Electrotechnical Commission (IEC) määrittelevät anonymisoinnin standardissa ISO/IEC 20889:2018:

“[P]rosessi, jolla henkilötieto [...] muunnetaan peruuttamattomasti tavalla, joka tekee tiedon kohteen tunnistamisen suoraan ja epäsuoraan mahdottomaksi, sekä rekisterinpitäjän toimesta yksin että yhteistyössä minkä tahansa kolmannen osapuolen kanssa.” (ISO/IEC 20889:2018, tutkielman kirjoittajien suomennos)

Euroopan Unionin tietosuoja-asetuksen GDPR:n periaatteessa 26 annetaan sisällöltään vastaava määritelmä anonyymeille tiedolle:

“[Tiedot], jotka eivät liity tunnistettuun tai tunnistettavissa olevaan luonnolliseen henkilöön, tai henkilö[tiedot], joiden tunnistettavuus on poistettu siten, ettei rekisteröidyn tunnistaminen ole tai ei ole enää mahdollista.” (EU-asetus 679/2016)

Itä-Suomen yliopiston tietosuojavastaava määrittelee anonymisoinnin seuraavasti: “Anonymisointi tarkoittaa henkilötietojen käsittelyä niin, että henkilöä ei enää voida tunnistaa niistä. Tunnistamisen täytyy kuitenkin estyä peruuttamattomasti ja siten, että kukaan ei voi enää muuttaa tietoja takaisin tunnistettaviksi, ja alkuperäiset tunnisteelliset tiedot on hävitetty. Anonymisoituja tietoja ei tulkita henkilötiedoiksi, ja niiden käsittelyyn ei sovelleta tietosuojalainsäädäntöä.” (Eronen, 2019)

Nämä määritelmät ovat ongelmallisia, koska niiden vaatimukset saattavat olla teoreettisesti mahdottomia. Tätä käsittelevät Rocher, Hendrickx ja Montjoye (2019). On myös huomattavaa, että jos oletetaan maailmankaikkeuden olevan kausaalisesti deterministinen, on täydellinen palauttamiskelvoton anonymisointi mahdotonta. Eli tuntemalla maailmankaikkeuden nykyinen tila voitaisiin johtaa myös maailmankaikkeuden anonymisointia edeltänyt tila. Tiedon teoreettista häviämättömyyttä testasivat muiden muassa Braunstein ja Pati (2007).

Anonymisaatio-termin laillista merkitystä koskeva problematiikka ei kuulu tämän tutkielman laajuuteen. Tästä johtuen tässä tutkielmassa käytetään seuraavaa määritelmää:

“Anonymisaatio on toimenpide tai joukko toimenpiteitä, joilla tieto muutetaan muotoon, josta yhdenkään yksittäisen ihmisen tunnistaminen ei ole realistisesti mahdollista ihmiskunnan tällä hetkellä tuntemalla teknologialla.” Huomionarvoista on, että tässäkin määritelmässä täytyy ottaa huomioon saatavilla olevan taustatiedon määrä. (Wang, Jia, Ke, 2017)

3 Kirjallisuuskatsaus

Koska tutkimuksen aiheena oli palvelupolkujen anonymisointi, pyrittiin kirjallisuuskatsauksella löytämään jo kehitettyjä käyttötarkoitukseen soveltuvia menetelmiä ja niiden vahvuuksia ja heikkouksia. Lisäksi pyrittiin muodostamaan käsitystä anonymisoinnista kokonaisuutena, jotta tutkielmassa kehitetyn menetelmän anonymisointivaihe saataisiin suoritettua luotettavasti ja hyvin mahdollisimman täydellisesti tyydyttäen eri organisaatioiden ja instituutioiden määritelmät anonymisoinnista kuten kuvattu yllä.

Kirjallisuuskatsauksessa hyödynnettiin Google Search -hakukonetta, Google Scholar -hakukonetta ja Jyväskylän yliopiston kirjaston JYKDOK-hakukonetta. Alla olevassa Taulukossa 1 luetellaan kirjallisuuskatsauksessa käytettyjä hakusanoja. Hakusanoja käytettiin yksin ja yhdistelminä.

Anonymisation	Anonymization
Attack	Care pathway
Clinical	Data
Deanonymization	Deanonymisation
EHR	Electronic health record
Pathway	Patient data
Patient record	Pseudonymization
Pseudonymisation	Process
Process anoymisation	Process anonymization
Risk	Process data

Taulukko 1. Käytettyjä hakusanoja.

Varsinaista tutkimusta itse sairaanhoidon palvelupolkujen anonymisoinnista löytyi jonkin verran. Lähteitä koskien palvelupolkujen anonymisoinnin menetelmiä oli vain vähän ja se

lienee aihe, jonka tarpeellisuutta ollaan vasta hahmottamassa. Yksittäisen aineiston anonymisoinnista sen sijaan löytyy paljon tutkimusta.

3.1 Yksittäisen aineiston anonymisointi

Tässä kappaleessa selvitämme yleisimpiä aineistojen anonymisointimenetelmiä. EU:n työryhmä "The Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data" on aiemmin linjannut, että yksikään yleinen anonymisointimenetelmä, jotka esitellään alla ei täytä kaikkia GDPR-asetuksen vaatimuksia. Työryhmän johtopäätökset esitellään taulukossa 2. (EU Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques)

	Onko yksilöinti riski?	**Onko linkittäminen riski?**	**Onko inferenssi riski?**
Pseudonymisointi	Kyllä	Kyllä	Kyllä
Kohinan lisääminen	Kyllä	Ehkä	Ehkä
Korvaaminen	Kyllä	Kyllä	Kyllä
Aggregaatio tai *k*-anonymiteetti	Ei	Kyllä	Kyllä
l-diversiteetti	Ei	Kyllä	Ehkä
Differentiaalinen anonymisointi	Ehkä	Ehkä	Ehkä
Hajautusalgoritmi tai tokenisaatio	Kyllä	Kyllä	Ehkä

Taulukko 2. Eri anonymisaatiotekniikkojen riskiarviointi GDPR-asetuksen vaatimusten kannalta EU:n yksilöiden suojaamisesta yksityistiedon osalta -työryhmän linjaamana (kirjoittajien käännös).

Tästä johtuen GDPR-asetuksen vaatiman anonymisoinnin toteuttaminen vaatii useamman anonymisointitekniikan yhdistämistä, jotta asetuksen asettamat vaatimukset yksityistiedon anonymisoinnille täyttyvät. *k*-anonymiteettia, *l*-diversiteettia ja *t*-läheisyyttä terveydenhoidon datan anonymisoinnissa ovat tutkineet esimerkiksi Rajendran, Jayabalan ja Rana. Erilaisten anonymisointitekniikoiden yhdistelyn vaikutusta tutkimustuloksiin on tutkinut Podlesny, Kayem, Meinel ja Jungmann (2019). GDPR-asetuksen vaikutuksista lääketieteellisen tutkimuksen laatuun on tehty useampi tutkimus. Koska GDPR estää datan tutkimuskäytön anonymisoimattomana ilman potilaan tai asiakkaan suostumusta, on Wierda ym. vedonnut parempien anonymisointitekniikoiden puolesta omassa tutkimuksessaan. (Rajendran, Jayabalan & Rana, 2017) (Clarke ym., 2019) (Negrouk & Lacombe, 2018) (Quinn, 2017) (Wierda ym., 2018)

3.1.1 *k*-anonymiteetti

Yksittäisten aineistojen anonymisointia voi toteuttaa kategorisoimalla tai sensuroimalla sarakkeita ja/tai rivejä. Näillä keinoilla saavutettua anonymiteettiä voidaan kuvata *k*-anonymiteetin ominaisuudella. *k* on suure, joka kertoo kuinka monesta *k*-1 muusta henkilöstä yksittäistä henkilöä ei voida tunnistaa. Tällöin mitä suurempi *k*:n arvo on, sitä varmempi anonymisointi on. Käytännössä tähän päästään sensuroimalla kenttiä jotka eivät ole oleellisia tutkimuksen kannalta esimerkiksi tarkka katuosoite ja kategorisoimalla tietoja kuten lajittemalla tarkat iät ikäluokkiin. Tämä varmistaa sen, että kaikki kentät säilyvät totuudenmukaisina toisin kuin lisätessä ääntä, tiivistäessä tai vaihtamalla arvoja aineistossa. (Bayardo & Agrawal, 2005, s. 1)

k-anonymiteetti kuitenkin kärsii kahdenlaisista hyökkäyksistä sen tarjoamaa anonymiteettiä vastaan: hyökkäys homogeenisyyden perusteella ja hyökkäys taustatiedon perusteella. Hyökkäys homogeenisyyden perusteella perustuu siihen, että aineistossa yhden kategorisoidun sarakkeen arvot ovat suurelta osin samoja, jolloin mahdollista arvata kentän todellinen arvo. Hyökkäys taustatiedon perusteella hyödyntää ulkopuolelta saatavan tiedon soveltamista *k*-anonymisoituun aineistoon. Esimerkiksi tieto jonkin väestöryhmän korkeammasta tai alemmasta riskistä sairastua johonkin tiettyyn sairauteen voi auttaa

rajaamaan teoreettisen samankaltaisten k-1 henkilön listaa lyhyemmäksi ja siten tunnistaa oikea henkilö aineistosta. (Aggarwal & Yu, 2008, s. 20-23)

Potilaan tunniste	**Osoite**	**Syntymäpäivä**	**Diagnoosi**
A11	Viivaväylä 2	12.12.2012	I12
B23	Neliökaari 4	09.09.1999	F99
C88	Kuutiokuja 6	17.07.1997	K77

Taulukko 3. Taulu potilastietoja ennen *k*-anonymisointia.

Potilaan tunniste	**Osoite**	**Syntymäpäivä**	**Diagnoosi**
A11	*	2012	I12
B23	*	1999	F99
C88	*	1997	K77

Taulukko 4. Taulukon 3 *k*-anonymisoinnin jälkeen. Aineistossa osoitteet on sensuroitu ja syntymäpäivät kategorisoitu vuoden mukaan.

3.1.2 *l*-diversiteetti

l-diversiteetti on anonymisointimenetelmä, joka pyrkii korjaamaan *k*-anonymisoinnin puutteita vaatimalla, että *k*-anonymisoidun aineiston kaikki mahdolliset arvot ovat "hyvin edustettuina" vähintään *l*:lä arvolla, jossa *l* on mielivaltainen kokonaisluku. *l*-diversiteetti siten vaatii myös arvoryhmien sisäistä arvojen jakautumista, joka ehkäisee *k*-anonymiteetissä ilmaantuvia homogeenisyyttä ja sen perusteella toteutettavia hyökkäyksiä. Tästä johtuen mitä suurempi *l*:n arvo on, niin sitä paremmin edustettuina jokainen arvo on oman ryhmänsä sisällä ja sitä parempi anonymisointi on. (Ninghui, Tiancheng & Venkatasubramanian, 2007, s. 2-4)

l-diversiteetin *l* muuttujan arvon vaikutusta dataan on yleisesti lähestytty kolmella tapaa. Selkeä *l*-diversiteetti, jossa *l* on suoraan vähimmäisedustuksen määrä jokaiselle arvolle sarakkeessa. Lisäksi on kehitetty rekursiivinen (*c, l*)-diversiteetti ja entrooppinen *l*-

diversiteetti, jotka rajaavat liian yleisiä ja liian harvinaisia arvoja. (Ninghui ym., 2007, s. 2-4) (Aggarwal & Yu, 2008, s. 26-27)

3.1.3 Selkeä *l*-diversiteetti

Selkeässä (distinct) *l*-diversiteetissä jokainen arvo arvoryhmän sisällä on edustettuna vähintään *l* arvolla. Tämä on määritelmistä yksinkertaisin, mutta on toimiva useimmissa tapauksissa. (Ninghui ym., 2007, s. 2-4) (Aggarwal & Yu, 2008, s. 26-27)

ID	**Kustannus**	**Ikä**
1	100 - 199	18 - 29
2	100 - 199	18 - 29
3	0 - 99	0 - 9
4	0 - 99	0 - 9

Taulukko 5. Taulu aineistosta, jossa kustannus- ja ikä-attribuutit ovat *l*=2-diversifioituneet.

3.1.4 Rekursiivinen *l*-diversiteetti

Rekursiivisen (*c, l*)-diversiteetin ehtona on, että yleisesti ilmenevät arvot eivät ilmene liian usein ja harvinaiset arvot liian harvoin. Eli matemaattisesti ilmaistuna, olkoon m arvojen määrä sarakkeessa ja r_i , $1 \leq i \leq m$, se arvojen määrä jossa i:ksi useimmin esiintyvä arvo esiintyy aineistossa A. Jos $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$, tällöin aineisto A täyttää rekursiivisesti (*c, l*)-diversiteetin. Rekursiivinen (*c, l*)-diversiteetti tarjoaa jonkin verran paremman suojan kuin selkeä *l*-diversiteetti. (Ninghui ym., 2007, s. 2-4) (Aggarwal & Yu, 2008, s. 26-27)

3.1.5 Entrooppinen *l*-diversiteetti

Entrooppisen *l*-diversiteetin mukaan aineistolla on entrooppinen *l*-diversiteeti kun jokaiselle yhtäläiselle luokalle A pätee S(A) ≥ log(*l*), eli:

$$S(A) = -\sum_{t \in T} p(S,t) \log_{10} p(S,t),$$

jossa:

S(A) l-diversiteetin entropia aineistossa *A*,

t on tärkeä tieto, joka kuuluu tärkeiden tietojen alueeseen *T* ja

p(S, t) on se osa aineistosta, joka sisältää tärkeän tiedon *t*.

Entrooppinen *l*-diversiteetti myös tarjoaa paremman suojan kuin selkeä *l*-diversiteetti. Vertailua rekursiivisen (*c*, *l*)-diversiteetin ja entrooppisen *l*-diversiteetin tehokkuuden välillä ei kirjallisuudesta löydetty. Riippumatta "hyvin edustetun" määritelmästä, *l*-diversiteetin toteuttaminen vaikeutuu huomattavasti jokaisen tärkeän sarakkeen myötä, koska ongelman ratkaisun ulottuvuudet kasvavat. (Ninghui ym., 2007, s. 2-4) (Aggarwal & Yu, 2008, s. 26-27)

3.1.6 *t*-läheisyys

Anonymiteettia voidaan mitata tiedon määrällä, jonka havainnoija saa aineistosta. Ennen aineiston saamista, havainnoijalla on jokin pohjakäsitys X aineiston tunnistavista tiedoista. Tutkittuaan aineistoa havainnoijan saaman tiedon määrä on erotus X:n ja uuden käsityksen Y välinen erotus. (Ninghui ym., 2007, s. 4-8)

t-läheisyyden lähestymistapa anonymisointiin jakaa aineiston sisältämän tiedon kahteen kategoriaan: tietoon koko väestöstä ja tietoon kohteena olevista henkilöistä. Tällöin myös käsitysten kehittyminen voidaan jakaa X:n muutoksena koko väestön tietojen jakauman perusteella käsitykseksi Y, joka kehittyy käsitykseksi Z tarkastellessa yksittäisten kohteena olevien henkilöiden tietojen jakaumaa. *l*-diversiteetti pyrkii vähentämään X:n ja Z:n välistä erotusta, mutta *t*-läheisyys pyrkii vähentämään Y:n ja Z:n välistä erotusta. Tällöin *t*-läheisyys ei rajoita tietoa koko väestöstä, mutta vähentää tietoa yksittäisistä henkilöistä,

joka on yleisesti parempi tutkittaessa väestötason ilmiöitä. *t*-läheisyys pyrkii ratkaisemaan *k*-anonymiteetin ja *l*-diversiteetin ongelmakohdat käsittelemällä aineiston dataa jakaumina. Valitun aineiston osan jakauma saa poiketa jakaumaltaan enimmillään *t* verran koko ominaisuuden jakaumasta.

Kahden todennäköisyysjakauman väliselle etäisyydelle löytyy useita määritelmiä, joista ainakin kolmea on käytetty *t*-läheisyyden kontekstissa kirjallisuudessa. (Ninghui ym., 2007, s. 4-8). Ninghui ym. (2007) antaavat seuraavat määritelmät jakaumien A = (a_1, a_2, ... a_n) ja B = (b1, b_2, ... b_m) etäisyyksille:

1. Tilastollinen etäisyys:

$$D(A,B) = \sum_{i=1}^{n} \frac{1}{2} |a_i - b_i|,$$

2. Kullback-Leibler etäisyys:

$$D(A,B) = \sum_{i=1}^{n} a_i \ln\left(\frac{a_i}{b_i}\right),$$

3. maansiirtäjän etäisyys, jossa jakaumat kuvastavat kasoja maata ja funktion ratkaisu on pienin mahdollinen työ, jolla yksi kasoista saadaan siirrettyä toiseen:

$$D(A,B) = W(A,B,F) = \sum_{i=1}^{k} \sum_{j=1}^{k} d_{ij} f_{ij},$$

jossa d_{ij} on etäisyys *A*:n alkion *i* ja *B*:n alkon *j* välinen etäisyys ja f_{ij} on siirretyn massan virtaus alkiosta *i* alkioon *j* pienimmällä mahdollisella työllä. Lisäksi funktiolle pätee:

$$f_{ij} \geq 0, 1 \leq i \leq n, \qquad 1 \leq j \leq n$$

$$p_i - \sum_{j=1}^{n} f_{ij} + \sum_{j=1}^{n} f_{ij} = q_i, \qquad 1 \le i \le n,$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1.$$

Nämä rajoitteet varmistavat, että jakauma A siirtyy jakaumaan B virtaumalla F. (Ninghui ym., 2007, s. 4-8)

3.1.7 Differentiaalinen anonymiteetti

Differentiaalinen anonymiteetti ottaa toisenlaisen lähestymistavan datan anonymisointiin. Siinä missä edellä mainitut *k*-anonymiteetti, *l*-diversiteeti ja *t*-läheisyys ovat datan funktioita, differentiaalinen anonymiteetti on anonymisoinnin mekanismin funktio. Kuitenkin, *t*-läheisyys voi tuottaa differentiaalisesti anonymisoitua dataa, jos tietyt reunaehdot täyttyvät, jonka Domingo-Ferrer ja Soria-Comas osoittavat. Tällöin sen teoriassa pitäisi toimia mille tahansa datalle samalla tavalla riippumatta datan laadusta. Differentiaalinen anonymiteetti pyrkii siihen, että sen avulla käsitellystä datasta olisi tilastollisesti epätodennäköistä pystyä sanomaan onko jonkun tietyn henkilön tietoja käytetty julkaistun tilaston luomisessa. Tätä tilastollista todennäköisyyttä kuvastaa anonymiteettiparametri ε. Tällöin ε-differentiaalinen anonymiteetti voidaan matemaattisesti määritellä seuraavasti: Olkoon A satunnaistettu algoritmi ja A_i A:n arvojoukko. A on ε-differentiaalisesti anonyymi, jos kaikille dataseteille D_1 ja D_2, joiden erottava tekijä on vain yksi alkio ja kaikille alajoukoille S arvojoukossa A_i pätee:

$$P[A(D_1) \in S] \le e^{\varepsilon} \times P[A(D_2) \in S],$$

jossa:

$\varepsilon \in \mathbb{R}_+$

ja todennäköisyys P otetaan A:n satunnaisuuden yli. (Domingo-Ferrer & Soria-Comas, 2015) (Dwork & Roth, 2014, s. 20)

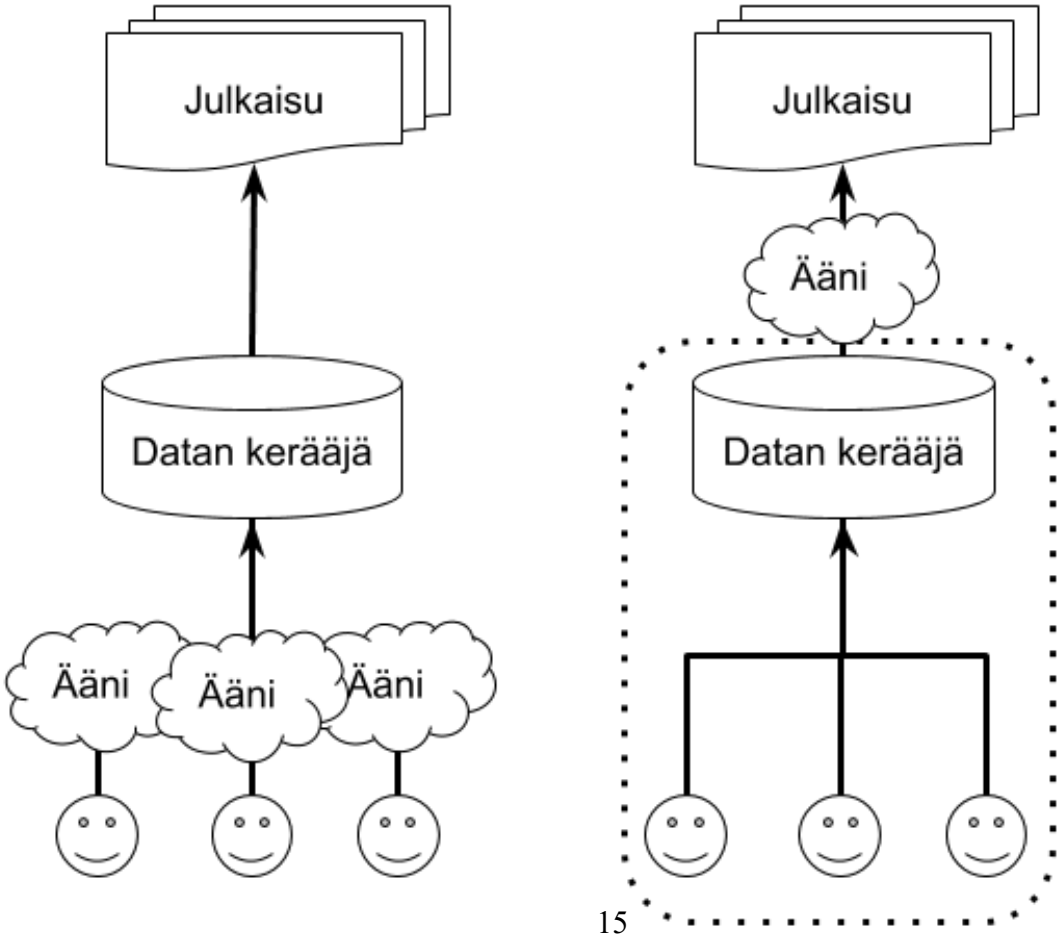
Koska tämä määritelmä on varsin tiukka on ε-differentiaalisesta anonymiteetistä kehitetty hieman joustavampi (ε, δ)-differentiaalinen anonymiteetti, joka on määritelmällisesti:

$$P[A(D_1) \in S] \leq e^{\varepsilon} \times P[A(D_2) \in S] + \delta,$$

jossa:

$\varepsilon \in \mathbb{R}_+$ ja relaksaatioparametri $\delta \in [0, 1]$. (Dwork, Roth, 2014, s. 20)

Toinen differentiaalisen anonymiteetin vahvuus on, että sitä voidaan käyttää luottamatta datan kerääjälle omaa raakadataa, vaan jokainen vastaus voidaan yksitellen anonymisoida, kuten esitetty alla.



Kuvio 1. Differentiaalisen anonymiteetin äänen voi lisätä kahdessa eri vaiheessa ennen data julkaisua.

Kuvion 1 kuvat esittävät, missä vaiheissa datan julkaisua differentiaalisen anonymiteetin kohinaa voidaan lisätä dataan. Ensimmäinen vaihtoehto on, että jokainen kyselyyn vastaaja lisää kohinaa omiin vastauksiinsa, jolloin vastaajien ei tarvitse luottaa datan kerääjään suojelemaan heidän anonymisoimatonta dataansa. Tämä metodi kuitenkin tuottaa enemmän kohinaa kuin toinen vaihtoehto. Toinen tapa on, että datan kerääjät lisäävät ääntä tuloksiin niitä aggregoidessa tai julkaistaessa eteenpäin. Tällöin kohinaa on lopputuloksessa vähemmän kuin kohinan yksilöllisessä lisäämisessä, mutta vastaajien pitää luottaa henkilötietoja sisältävä datansa datan kerääjän käyttöön. (Kairouz, Oh & Viswanath, 2014)

3.1.8 Laplacen mekanismi

Yksi yleisesti käytetty tapa, jolla ε-differentiaalinen anonymiteetti voidaan toteuttaa on Laplacen mekanismi, jossa dataan lisätään Laplacen jakauman perusteella tilastollista melua kunnes ε-differentiaalisen anonymiteetin vaatimus on täyttynyt. Laplacen todennäköisyystiheysjakauma, joka tunnetaan myös kaksoiseksponentiaalijakaumana voidaan määritellä:

$$f(\mu, b) = \frac{1}{2b} \exp\left(-\frac{\mu}{b}\right),$$

jossa:

μ = sijaintiparametri ja myös jakauman keskiarvo, mediaani ja moodi,

b = skaalaparametri, joka määrittää jakauman leveyden, tällöin pätee b > 0.

Käytetty Laplacen melu voidaan määrittää täyttävän ε-differentiaalisen anonymiteetin vaatimuksen kun:

$$\Delta f = max|f(x) - f(y)|,$$

jossa:

x = anonymisoitu data-aineisto,

y = alkuperäinen data-aineisto ja

Δf = funktion herkkyys, joka johdetaan x:n ja y:n yksittäisten elementtien suurimmalla erotuksella.

Funktion herkkyyden perusteella voidaan määrittää mekanismi M, jonka kautta voidaan julkaista funktion ftietoja ε-differentiaalisesti anonymisoituna:

$$M_{Laplace}(x, f, \varepsilon) = f(x) + Laplace\left(\mu = 0, b = \frac{\Delta f}{\varepsilon}\right),$$

jossa:

x = alkuperäinen data-aineisto,

f = alkuperäisen data-aineiston funktio ja

ε = ε-differentiaalisen anonymisoinnin aste.

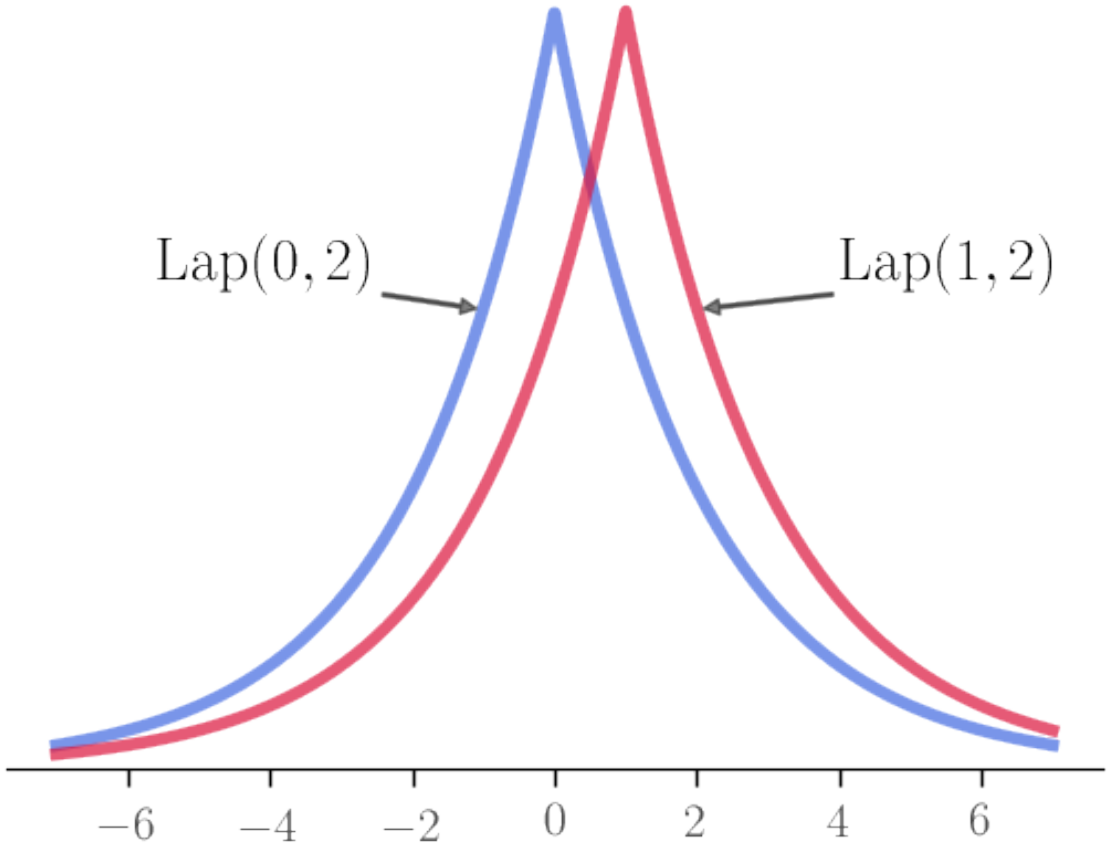
Anonymisointi on saavuttanut tarvittavan melun määrän, kun seuraava ehto on täyttynyt, eli kun tuotettu jakauman $M_{Laplace}(x, f, \varepsilon)$ on lähellä alkuperäistä jakaumaa$M_{Laplace}(y, f, \varepsilon)$ joka kohdassa:

$$\frac{Pr\left(M_{Laplace}(x, f, \varepsilon) = z\right)}{Pr\left(M_{Laplace}(y, f, \varepsilon) = z\right)} = \frac{Pr\left(f(x) + Laplace\left(0, \frac{\Delta f}{\varepsilon}\right) = z\right)}{Pr\left(f(y) + Laplace\left(0, \frac{\Delta f}{\varepsilon}\right) = z\right)}$$

$$\frac{Pr\left(M_{Laplace}(x, f, \varepsilon) = z\right)}{Pr\left(M_{Laplace}(y, f, \varepsilon) = z\right)} = \frac{\frac{1}{2b} exp\left(-\frac{|z - f(x)|}{b}\right)}{\frac{1}{2b} exp\left(-\frac{|z - f(y)|}{b}\right)}$$

$$\frac{Pr\left(M_{Laplace}(x, f, \varepsilon) = z\right)}{Pr\left(M_{Laplace}(y, f, \varepsilon) = z\right)} \leq exp\left(\frac{\Delta f}{b}\right) = exp(\varepsilon)$$

Epäyhtälö lausekkeesta muodostuu kolmioepäyhtälöstä ja funktion herkkyyden ehdosta. (Sarathy, Muralidhar, 2011, s. 6-7) (Dwork, Roth, 2014, s. 20)

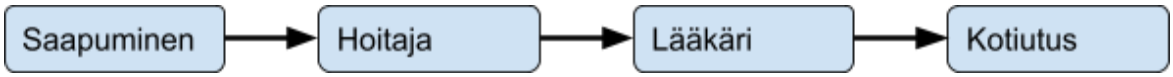


Kuvio 2. Laplacen jakaumat Lap(μ=0, b=2) ja Lap(μ=1, b=2), jotka tuottavat 0,5-differentiaalisen anonymiteetin funktioille, joiden herkkyys on 1ja keskiarvo vastaavasti 0 ja 1.

Esimerkki jakaumista (0, 5)-differentiaalisen anonymiteetin tuottamisesta funktiolle, jonka herkkyys on 1 esitetään kuviossa 2. Koska anonymiteettimuuttuja ε vaikuttaa kokoparametriin b, on jakauma sitä leveämpi mitä pienempi ε:n arvo on, jolloin myös tuotettu melu saa todennäköisemmin suurempia arvoja.

4 Palvelupolku

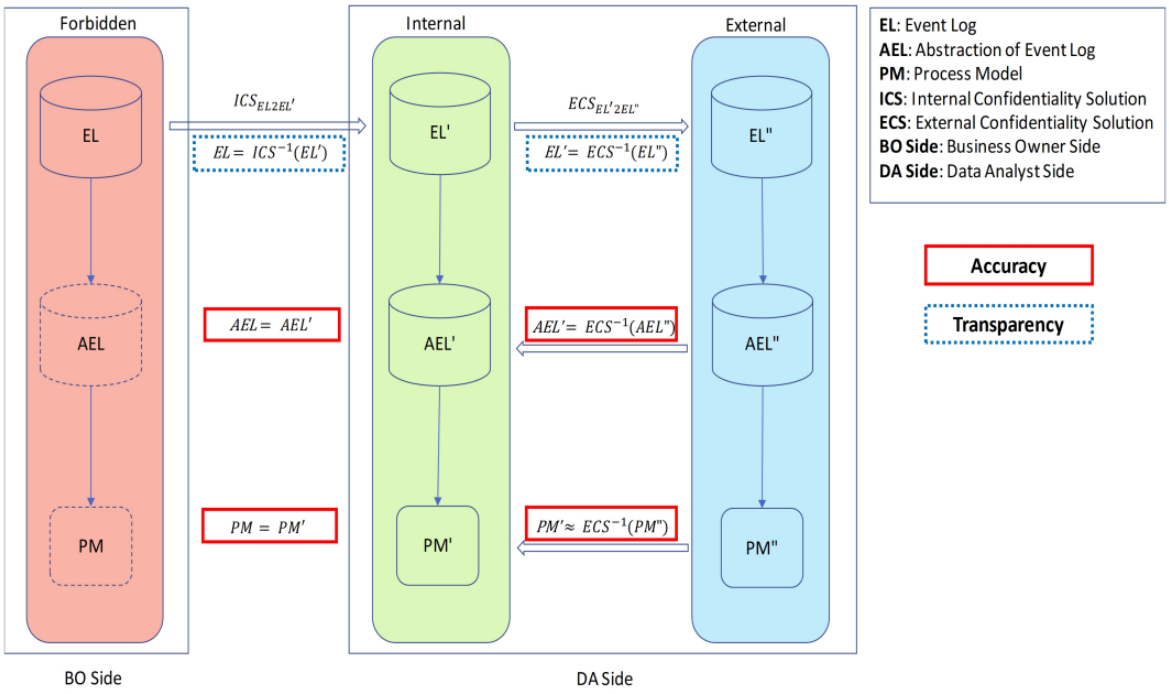
Palvelupolku on potilaan yksittäisistä hoitotapahtumista koostuva ketju tapahtumia. Yksinkertainen palvelupolku on esitetty kuviossa 3. Palvelupolut voivat perustua malleihin, joita luodaan ennalta sairaaloissa tiettyjä sairauksia ja vaivoja varten. Tämä yhtenäistää käytäntöjä potilaiden välillä ja määrittää seuraavia toimenpiteitä ennalta, säästäen resursseja ja aikaa. Kuitenkin jokaisen potilaan palvelupolku voi poiketa ennaltamääritetyistä malleista ihmisten, sairauksien ja vaivojen yksilöllisyyden takia. Tästä syystä palvelupolku voi olla hyvinkin yksilöivä datan palanen esimerkiksi harvinaisen sairauden, komplikaation tai hoidon ajankohdan takia (Linsman, Rotter, James, Snow & Willis, 2010).



Kuvio 3. Yksinkertainen palvelupolku.

4.1 Palvelupolkujen anonymisointi ja sen menetelmät

Pika, Wynn, Budino, ter Hofstede ja van der Aalst (2019) käsittelevät monipuolisesti anonymisaatiota prosessilouhinnan kontekstissa. Tämän tutkielman kannalta erityisen relevantti on heidän artikkelinsa kappale 4.2 Anonymising Atypical Process Behaviour, joka käsittelee muiden asioiden muassa myös harvinaisten prosessivarianttien (eli “infrequent activity sequences”) anonymisointia. He ottavat esille kaksi mahdollista menetelmää. Nämä ovat Rafiein, von Waldthausenin ja van der Aalstin (2018) luottamuksellisuuskehikko (eli “confidentiality framework”) ja Fahrenkrog-Petersenin, van der Aan ja Weidlichin (2019) PRETSA. Luottamuksellisuuskehikko on esitetty kuviossa 4 ja PRETSA-menetelmän vertailua on esitetty kuviossa 5.



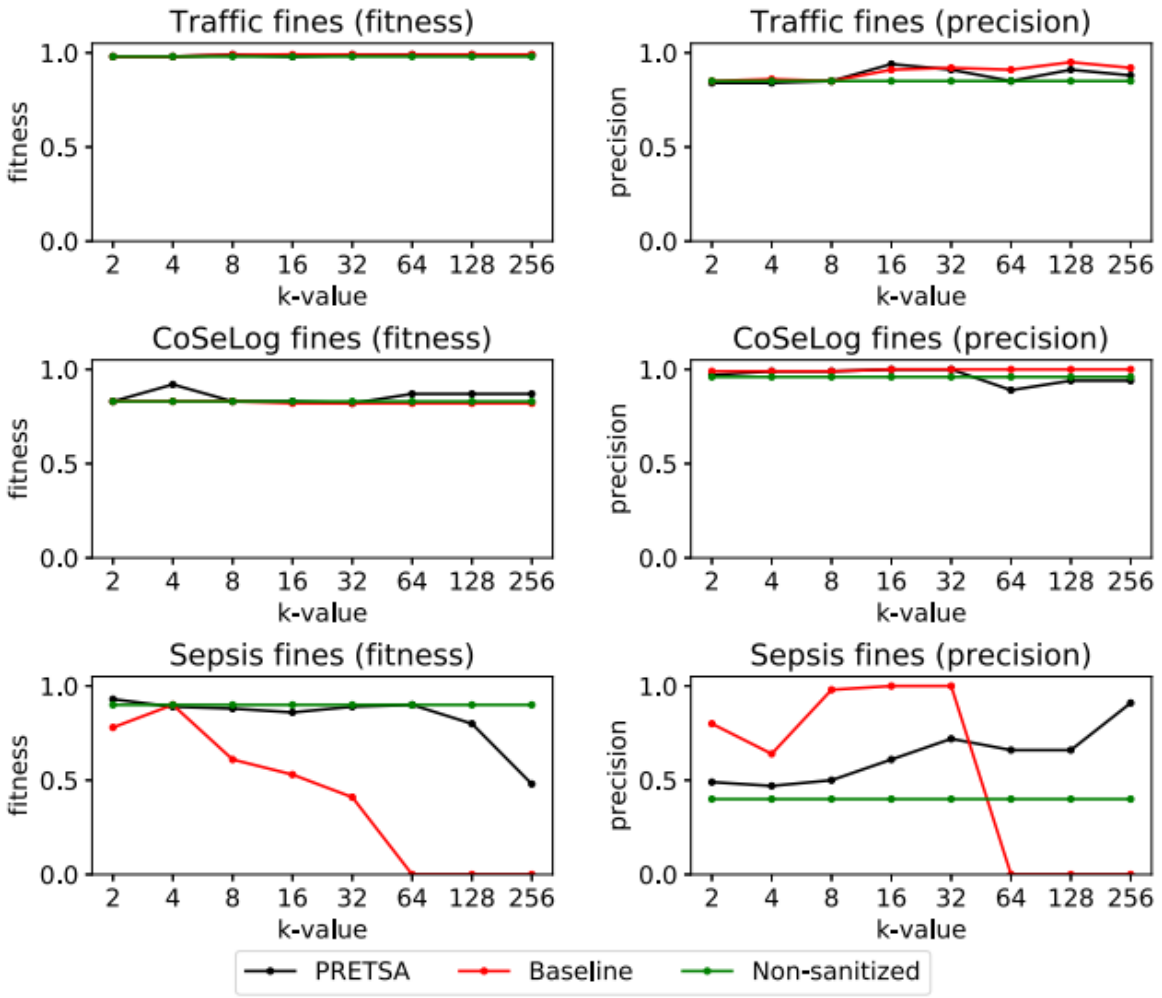
Kuvio 4. Luottamuksellisuuskehikon kolme kerrosta (Refiei ym. 2018).

Varsinaisen palvelupolkujen anonymisoinnin Rafiei ym. (2018) toteuttavat muuttamalla aikaleimat suhteellisiksi, muotoon, jossa niistä ei enää voi päätellä tapahtumien keskinäistä järjestystä. Tämän sijaan he lisäävät tapahtumiin kryptatun linkin edelliseen tapahtumaan ja tiedon siitä, mitä aktiviteettia edellinen tapahtuma edusti. Lisäksi entiteetin ja/tai prosessi-instanssin tunnisteet poistetaan datasta. Harvinaisten varianttien anonymisointi tehdään menetelmässä poistamalla ne datasta.

Kuten Pika ym. (2020) huomauttavat, Rafiein ym. (2018) menetelmällä anonymisoituun dataan kohdistuva prosessinlouhinta vaatisi erikoistuneet työkalut. Lisäksi, kuten Pika ym. (2020) mainitsevat, merkittävä osa etenkin terveydenhuollon prosessivarianteista voi olla uniikkeja, joten ne menetettäisiin Rafiein ym. (2018) menetelmää käytettäessä kokonaan. Tämä olisi erityisen totta datoille, joissa ei ole prosessi-instanssi-kohtaista tunnistetta, vaan pelkkä entiteettikohtainen tunniste.

Fahrenkrog-Petersenin ym. (2019) PRETSA on menetelmä, jossa prosessin variantit esitetään puurakenteena. He poistavat solmut, jotka eivät täytä asetettua *k*-anonymiteetti-

tai *t*-läheisyys-vaatimusta.



Kuvio 5. PRETSA-menetelmällä anonymisoidun datan analyysi verrattuna alkeellisemmilla menetelmillä anonymisoituun dataan ja anonymisoimattomaan dataan (Fahrenkrog-Petersen ym. 2019).

Tässä tutkielmassa käytetyt anonymisointimenetelmät ovat verrattavissa Fahrenkrog-Petersenin ym. (2019) käyttämiin alkeellisiin menetelmiin, mutta esittelemämme polkuattribuutti-menetelmä on yhteen sopiva muidenkin anonymisointi menetelmien kanssa.

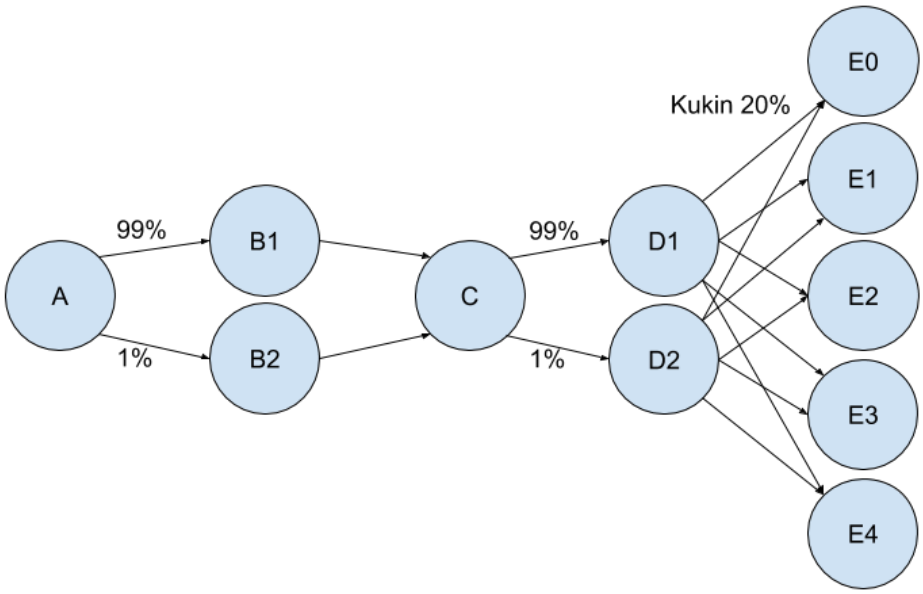
Kuten Pika ym. (2020) ottavat esille myös PRETSA:lle on yhteistä se, että se poistaa uniikkeja prosessivariantteja, mikä on ongelmallista terveydenhuollossa ja sen kaltaisissa konteksteissa. Kuten Rafiein ym. (2018) menetelmässä myös PRETSA:ssa tämä ongelma kosostuu, mikäli data sisältää vain entiteettien tunnisteet.

Teimme sen havainnon koskien kuvattuja menetelmiä, että niiden hyödyntäminen vaatii jo itsessään varsin prosessilouhinnan kaltaista analyysiä. Prosessilouhinta on käytännössä prosessi-instanssien tapahtumien löytämistä ja järjestämistä prosessin mukaisesti. Jotta datan tapahtumariveille saadaan merkittyä myös viite prosessi-instanssien seuraaviin tapahtumiin, kuten Rafiein ym. (2018) menetelmä edellyttää, pitää prosessi-instanssin tapahtumien jo olla tiedossa. Vastaavasti prosessivarianttien todennäköisyyksien kuvaaminen PRETSA-tyyppisessä puu muodossa vaatisi sitä, että kunkin variantin instanssit on jo tunnistettu datasta.

5 Menetelmä

Tutkimuksen tavoitteena on luoda menetelmä, jolla voidaan anonymisoida yksittäisten tapahtuma-tietojen lisäksi myös prosessin kulku eri tapahtumien välillä. Tätä tarkoitusta varten luodaan terveydenhuollon tapahtumadataa jäljittelevä pseudodata-tiedosto. Datassa prosesseina toimivat siis geneeriset pseudohoitopolut. Pseudodatan luomista varten kirjoitetaan datan generointi -ohjelma. Luotu data ja generaattori liitteestä A.

Generoitava data mallintaa seuraavan prosessin läpimenodataa:



Kuvio 6. Prosessi generoidussa datassa.

Dataan generoidaan 100 000 (satatuhatta) asiakasta, joista kullakin on prosessin mukaisesti viisi tapahtumaa. Erilaisia prosessivariantteja syntyy kuvion 6 mukaisten solmujen yhdistelminä 20 erilaista hoitopolkua prosessin lävitse. Datan yleisimmät variantit ovat

muotoa {A->B1->C->D1->E#}, jotka kattavat 98,01% kaikista prosessi-instansseista. Seuraavaksi yleisimmät variantit ovat muotoa {A->B1->C->D2->E#} tai {A->B2->C->D1->E#]. Kumpikin näistä muodoista kattaa prosessi-instansseista 0,99%. Harvinaisimmat variantit ovat muotoa {A->B2->C->D2->E#} ja esiintyvät vain 0,01%:lla asiakkaista, eli satatuhatta asiakasta sisältävässä aineistossa odotusarvoisesti yhteensä kymmenen kertaa. Koska kunkin eri E-vaiheen tapahtuman todennäköisyys on 20%, on näistä kunkin yksittäisen variantin odotettu absoluuttinen esiintymismäärä sadantuhannen prosessi-instanssin aineistossa kaksi.

Variantit {A->B2->C->D2->E#} ovat siis hyvin harvinaisia ja voivat toimia tunnistavina tekijöinä aineistossa. Tutkielma keskittyy niiden anonymisointiin. Harvinaisia variantteja esiintyy genoroidussa datassa taulukon 6 mukaisesti.

Variantti	**Absoluuttinen esiintymisfrekvenssi**
{A->B2->C->D2->E0}	1
{A->B2->C->D2->E1}	1
{A->B2->C->D2->E2}	1
{A->B2->C->D2->E3}	2
{A->B2->C->D2->E4}	6

Taulukko 6. Harvinaisten varianttien esiintymisfrekvenssi aineistossa.

5.1 Polkuattribuutti

Datan normaali muoto, ja myös sen normaalimuoto tietokannassa, vastaisi taulukon 7 sisältöä.

Asiakas	**Aktiviteetti**	**Aloitusaika**	**Lopetusaika**
1	A	2019-02-01 00:00:00	2019-02-01 01:01:00
1	B1	2019-02-01 02:03:00	2019-02-01 03:06:00

1	C	2019-02-01 04:10:00	2019-02-01 05:15:00
1	D1	2019-02-01 06:21:00	2019-02-01 07:28:00
2	A	2019-06-02 00:00:00	2019-06-02 01:00:00
...	...	...	...

Taulukko 7. Datan normaali muoto aineistossa.

Kyseisessä muodossa kukin datan rivi kuvaa yhtä tapahtumaa. Ensimmäinen sarake kertoo mitä asiakasta kyseinen tapahtuma koskee ja toinen sarake mikä aktiviteetti on suoritettu. Kolmas ja neljäs sarake kertovat tapahtuman alkamis- ja päättymisajat.

Käsittelyn helpottamiseksi ja toimenpiteiden selkeyttämiseksi tutkimuksessa muunnetaan yllä oleva muoto abstraktimmaksi. Abstraktimpaa muotoa kutsutaan tutkielmassa polkuattribuuttimuodoksi. Datan polkuattribuuttimuotoon muuttaminen koostuu kolmesta vaiheesta, jotka on kuvattu alla.

5.2 Vaihe 1, polkuaattribuutin luominen

Lisätään asiakkaan kullekin riville uusi attribuutti, joka kertoo asiakkaan palvelupolun ja tapahtumien sekä siirtymien kestot. Muoto, jota tutkielmassa käytetään attribuutin esittämiseen on:

$$[1.\,aktiviteetti]([kesto]):([siirtymän\ kesto]):[2.\,aktiviteetti]([kesto]):\\([siirtymän\ kesto]):\ldots:[viimeinen\ aktiviteetti]([kesto]).$$

Kestot esitetään kokonaislukuina, joka edustaa jotain aikayksikköä. Aikayksikön tulee olla yhtenäinen kaikille polkuattribuuteille; tutkielmassa käytetään minuutteja.

Esimerkiksi yllä olevassa taulukossa 7 esiintyvän asiakkaan 1 polkuattribuutti olisi “A(61):(62):B1(63):(64):C(65):(66):D1(67)”. Kaikki asiakkaan 1 rivit datassa polkuattribuuteilla täydennettynä näyttäisivät taulukon 8 kaltaisilta.

Asiakas	Aktiviteetti	Aloitusaika	Lopetusaika	Polku
1	A	2019-02-01 00:00:00	2019-02-01 01:01:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)
1	B1	2019-02-01 02:03:00	2019-02-01 03:06:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)
1	C	2019-02-01 04:10:00	2019-02-01 05:15:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)
1	D1	2019-02-01 06:21:00	2019-02-01 07:28:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)

Taulukko 8. Esimerkki-asiakkaan 1 tapahtumat polkuattribuutilla.

Tutkimuksessa käytettävä datageneraattori tuottaa datan tässä muodossa turhien työvaiheiden välttämiseksi.

5.3 Vaihe 2, sarakkeiden poisto

Poistetaan datasta kaikki sarakkeet, paitsi asiakastunnus (esimerkissä “Asiakas”), aloitusaikaleima (esimerkissä “Alk. aika”) ja polkuattribuutti (esimerkissä “Polku”). Tämä voidaan tehdä käytetylle datalle tietoja menettämättä, koska aktiviteetit löytyvät myös polkuattribuutista ja loppuaikaleimat saadaan johdettua kestoista.

Poiston jälkeen esimerkki-asiakkaan 1 datarivit näyttäisivät taulukon 9 kaltaiselta.

Asiakas	Aloitusaika	Polku
1	2019-02-01 00:00:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)
1	2019-02-01 02:03:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)
1	2019-02-01 04:10:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)
1	2019-02-01 06:21:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)

Taulukko 9. Esimerkki-asiakkaan 1 tapahtumat polkuattribuutilla.

On huomionarvoista, että vaikka tässä tutkimuksessa ei datan sisällön takia menetetä tietoja sarakkeita poistaessa, ei tämä päde kaikkiin mahdollisiin data-aineistoihin. Todelliset aineistot voivat sisältää tapahtumakohtaista tietoa, kuten esimerkiksi tapahtuman kustannuksen tai suorituspaikan tai asiakaskohtaista tietoa, kuten kotipaikkakunnan tai iän. Tämän kaltaisten attribuuttien säilyttäminen anonymisointiprosessin yli ei sisälly tämän tutkielman laajuuteen.

5.4 Vaihe 3, rivien poisto

Ensimmäistä tapahtumaa seuraavien tapahtumien alkuaikaleimat voidaan johtaa ensimmäisen tapahtuman alkuaikaleimasta ja polkuattribuutista. Näiden tapahtumien rivit voidaan siis poistaa aineistosta menettämättä tietoja.

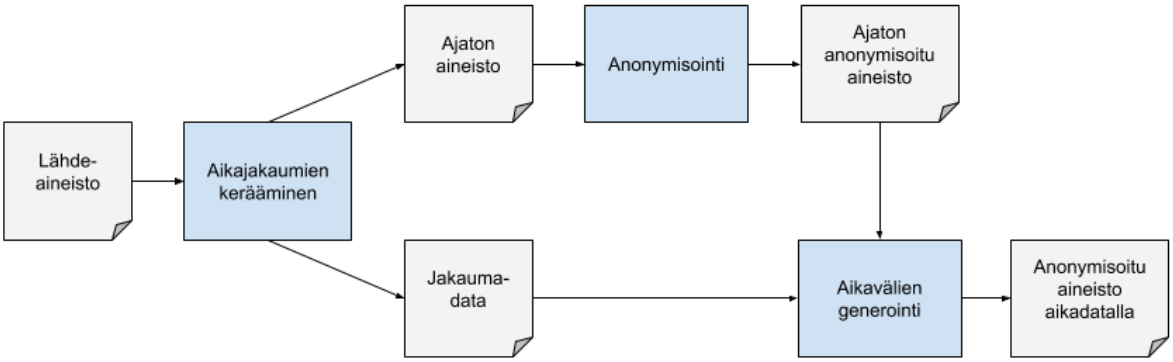
Tuloksena saadaan polkuattribuuttimuodossa oleva aineisto. Tässä muodossa kullakin asiakkaalla on yksi rivi, joka sisältää asiakkaan tunnisteen, asiakkaan koko prosessi-instanssin alkamisajan, sekä polkuattribuutin. Esimerkki-asiakkaan 1 rivi olisi taulukon 10 kaltainen.

Asiakas	Aloitusaika	Polku
1	2019-02-01 00:00:00	A(61):(62):B1(63):(64):C(65):(66):D1(67)

Taulukko 10. Esimerkki-asiakkaan 1 tapahtumat polkuattribuutilla.

5.5 Anonymisointiprosessi

Tutkittava anonymisointiprosessin lähdeaineisto on yllä kuvatussa polkuattribuuttimuodossa. Vastaava prosessi olisi mahdollista suorittaa myös aineiston ollessa normaalissa muodossa, mutta vaatisi erikoistyökalujen luomista. Polkuattribuuttimuoto mahdollistaa olemassa olevien anonymisaatiotyökalujen käytön. Anonymisointiprosessin rakenne on esitetty kuviossa 7.



Kuvio 7. Anonysointiprosessin rakenne.

5.6 Aikajakaumien kerääminen

Aikajakaumilla tarkoitetaan tapahtumien kestojen jakaumia ja tapahtumien välisten siirtymien kestojen jakaumia. Aikajakaumien keräämiseen käytetään tutkielmaa varten luotua DistributionCollector-ohjelmaa. Aikajakaumien keräämisen lisäksi ohjelma tuottaa aineistosta version, jonka polkuattribuuteista kestotiedot on poistettu.

Ohjelma kerää aikajakaumat prosessivarianttikohtaisesti aineiston polkuattribuuttimuodosta. Jakaumatiedot kerätään varianttikohtaisesti, koska eri aktiviteettien kestot saattavat vaihdella suurestikin eri varianttikontekstien välillä. Esimerkiksi terveydenhuollon lähdeaineistosta mahdollisesti löytyvän aktiviteetin “Leikkaus” kesto voi vaihdella paljonkin, riippuen siitä onko kyseessä esimerkiksi aivokirurgia vai luomen poisto.

Tutkimuksessa käytetty versio DistributionCollector-ohjelmasta kerää varianttien aikatiedot diskreeteiksi empiirisiksi jakaumiksi. Pienillä aineistoilla, kuten harvinaisten

varianttien aikaleimoilla, diskreettien empiiristen jakaumien tuottamat arvot voivat olla liian homogeenisia. Tähän problematiikkaan palataan alla alikappaleessa “Aikavälien generointi”.

On mahdollista, että aikaan saataisiin tarkempia tuloksia käyttämällä jakaumien sovittamista (eng. distribution fitting). Tähän ei kuitenkaan tutkimuksessa päädytty kahdesta syystä. Ensinnä jakaumien sovittaminen on työlästä suhteessa tutkimukselle tuotettuun lisäarvoon. Toiseksi tutkielma keskittyy asiakkaiden tunnistamisen mahdollistaviin harvinaisiin variantteihin, joiden esiintymismäärä on niin pieni, etteivät ne tarjoa hyviä edellytyksiä jakaumien sovittamiseen.

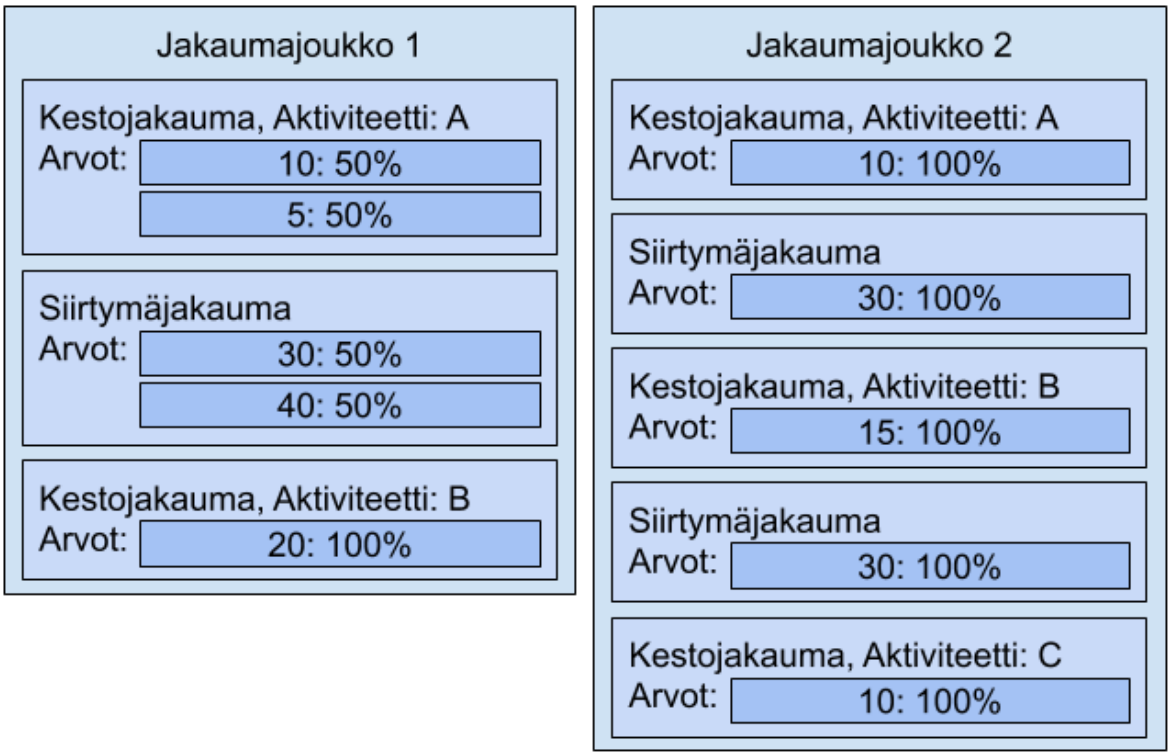
Aikajakaumatietojen keräämisen lisäksi tässä anonymisointiprosessin vaiheessa luodaan aineistosta versio, joka ei sisällä aikavälitietoja. Käytännössä tämä tarkoittaa, että tuotetussa aineistossa asiakkaiden polkuattribuutti on se prosessivariantti, jonka instanssin asiakas on toteuttanut.

Otetaan esimerkiksi aikajakaumien keräämisestä lähdeaineisto, joka on esitetty taulukossa 11.

Asiakas	**Aloitusaika**	**Polku**
1	2019-01-01 01:00:00	A(10):(30):B(20)
2	2019-01-02 02:00:00	A(5):(40):B(20)
3	2019-01-03 03:00:00	A(10):(30):B(15):(30):C(10)

Taulukko 11. Aikajakaumien keräämisen esimerkkiaineisto.

Aineiston pohjalta syntyisi kaksi varianttikohtaista jakaumajoukkoa, jotka on esitetty kuvi-ossa 8.



Kuvio 8. Esimerkkiaineistosta kerätyt aikajakaumajoukot.

Jakaumajoukkojen ohella syntyisi aineisto ilman aikavälitietoja, joka on esitetty taulukossa 12.

Asiakas	**Aloitusaika**	**Polku**
1	2019-01-01 01:00:00	A:B
2	2019-01-02 02:00:00	A:B
3	2019-01-03 03:00:00	A:C

Taulukko 12. Aikajakaumien keräämisen esimerkkiaineisto.

5.7 Anonymisointi

Aineiston palvelupolun solmujen välisestä ajasta riippumattomat muuttujat, eli alkuaikaleima ja polun solmujen rakenne voidaan esimerkiksi *k*-anonymisoida tai *l*-diversifioida riippuen aineiston koostumuksesta ja määrästä. Uniikkien polkujen anonymisointi vaatisi muun datan anonymisointia samalle tasolle, joka voi johtaa aineiston

käyttökelvottomaan tarkkuuteen. On kuitenkin otettava huomioon, että valtavirrasta poikkeavien ja uniikkien, mutta validien alkioiden anonymisointi siten, että aineisto säilyttää hyödyllisen määrän tarkkuutta, voi olla mahdotonta (Sarathy, Muralidhar, 2011, s. 6-7).

Varsinainen anonymisointi kohdistuu luotuun polkuattribuuttisarakkeeseen. Anonymisointivaiheen tarkoituksena on vähentää sen sisältämiä tunnistavia arvoja. Polkuattribuuttimuotoisen datan tunnistesarakkeen kuuluu olla uniikki jälkikäsittelyä varten, joten sen anonymisointi ei ole mielekästä. Alkuaikaleima-sarakkeen anonymisointi voi olla tapauskohtaisesti järkevää. Esimerkiksi, mikäli dataa anonymisoidaan kuukausitarkastelua varten, kellonaikatiedot eivät ole relevantteja ja ne voidaan poistaa. Tämän tutkielman kannalta aloitusaikaleiman anonymisointi ei ole relevanttia.

Tutkimuksessa käytetyn polkuattribuutti-lähestymistavan kannalta varsinaiseen anonymisointiin käytetyllä työkalulla ei ole merkitystä. Tässä tutkimuksessa anonymisointi-vaiheessa päädyttiin käyttämään ARX Data Anonymization Tool -ohjelmaa (Prasser, Eicher, Spengler, Bild & Kuhn, 2020).

5.8 Aikavälien generointi

Jotta anonymisoitu aineisto saadaan samaan muotoon kuin lähdeaineisto, on siihen lisättävä aikavälitiedot. Tähän voidaan käyttää aiemmin kerättyjä aikajakaumia. Jakaumien generointi voidaan myös hoitaa DistributionGenerator-ohjelmalla.

DistributionGenerator sisältää toiminnot arvojen generointiin suoraan DistributionCollector-ohjelmalla kerätyistä empiirisistä jakaumista sekä mahdollisuuden lisätä generoituihin arvoihin Laplace-jakaumalla tuotettua satunnaisuutta. Satunnaisuuden tuottavan Laplace-jakauman sijaintiattribuutti on nolla ja skaala-attribuuttina käytetään joko jakauman keskihajontaa tai käyttäjän määrittämää minimiarvoa, mikäli se on suurempi.

DistributionGenerator vertaa aiemmin generoitujen jakaumajoukkojen aktiviteettien järjestystä anonymisoidun aineiston polkuattribuuttiin. Löydettyään polkuattibuuttia

vastaavan jakaumajoukon se generoi polun aktiviteeteille kestot ja niiden väleille siirtymäajat.

Oletetaan esimerkiksi, että anonymisointivaiheen läpi ollaan viety aikajakaumien kerääminen -vaiheessa tuotettu aikavälitiedoton aineisto, joka on esitetty taulukossa 13.

Asiakas	**Alkuaikaleima**	**Polku**
1	2019-01-01 01:00:00	A:B
2	2019-01-02 02:00:00	A:B
3	2019-01-03 03:00:00	A:C

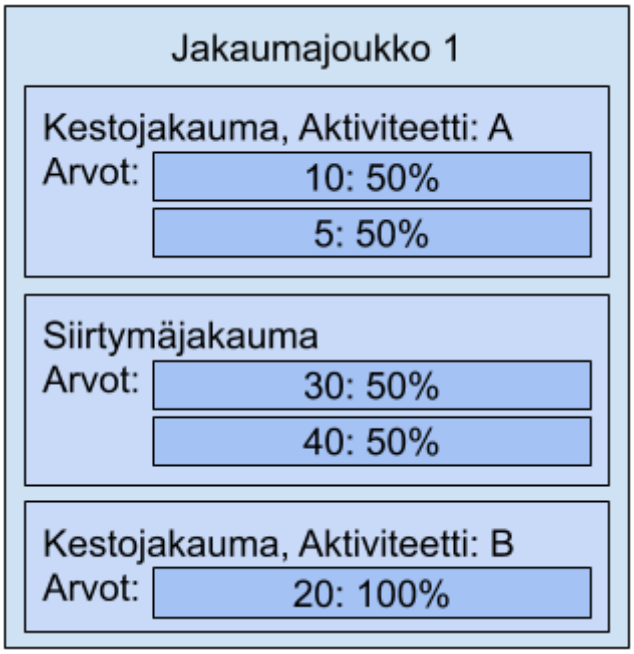
Taulukko 13. Esimerkkiaineisto ilman aikavälitietoja.

Oletetaan, että anonymisointi-vaiheessa ollaan asiakkaan 3 anonymisoimiseksi muutettu tämän polkuattribuutti vastaamaan yleisempää polkuattribuuttia, joka on esitetty taulukossa 14.

Asiakas	**Aloitusaika**	**Polku**
1	2019-01	A:*
2	2019-01	A:*
3	2019-01	A:*

Taulukko 14. (k=3)-anonymisoitu esimerkkiaineisto ilman aikavälitietoja.

DistributionGenerator tunnistaisi kunkin polun jakaumajoukon 1 mukaiseksi, joka esitetty kuviossa 9.



Kuvio 9. Jakaumajoukko 1

Kuhunkin anonymisoidun datan palvelupolkuun lisättäisiin jakaumajoukon 1 pohjalta generoidut aikavälitiedot, joka esitetty taulukossa 15. Tässä esimerkissä ei ole käytetty Laplace-jakaumaan pohjautuvaaa satunnaisuutta.

Asiakas	**Aloitusaika**	**Polku**
1	2019-01-01 01:00:00	A(5):(30):B(20)
2	2019-01-02 02:00:00	A(10):(30):B(20)
3	2019-01-03 03:00:00	A(5):(40):B(20)

Taulukko 15. Anonymisoitu esimerkkiaineisto generoiduilla aikavälitiedoilla.

6 Koe 1: Polkuattribuutin käyttäminen anonymisoinnissa

Kokeessa 1 pyritään anonymisoimaan generoitu pseudodata käyttämällä polkuattribuuttimuotoa. Polkuattribuutti-muodossa olevan pseudodatan kestojakaumat kerätään, polkuattribuutti-sarake anonymisoidaan ja kerättyjen kestojakaumien pohjalta generoidaan uudet kestot anonymisoiduille riveille.

6.1 Kokeen eteneminen

Tutkimuksen suoraviivaistamiseksi lähdeaineisto generoitiin suoraan polkuattribuuttimuotoon, joka löytyy liitteestä A nimellä ”data10000”. Generoitu aineisto ajettiin DistributionCollector-ohjelman lävitse, mikä tuotti aikaleimattoman version aineistosta sekä erillisen jakaumadatan, jotka löytyvät vastaavasti liitteestä A nimillä ”data10000_timeless” ja ”data10000_distributions”.

Aikaleimaton aineisto anonymisoitiin käyttämällä ARX Data Anonymization Tool -ohjelmaa. Anonymisointiin kokeiltiin automaattista relaksoitua (ε=2, δ=1x10^-6)-differentiaalista anonymisointia, automaattista (k=2)-anonymisointia ja lopuksi karkeaa winsorizing-tyyppistä sensurointia. (Prasser, Eicher, Spengler, Bild & Kuhn, 2020) (Tukey, 1962)

6.1.1 Anonymisointi 1: Relaksoitu differentiaalinen anonymisointi

Relaksoitu (ε=2, δ=1x10^-6)-differentiaalinen anonymisointi poisti datasta kaiken tunnistavaksi tai pseudotunnistavaksi luokitellun tiedon. Tämä ei ollut hyväksyttävää, koska poistetun tiedon määrä oli merkittävä osa aineistosta.

6.1.2 Anonymisointi 2: Automaattinen (*k*=2)-anonymisointi

Automaattinen (*k*=2)-anonymisointi (right-to-left) totesi tunnistavaksi tekijäksi polkuattribuutin viimeisen merkin ja korvasi E-vaiheen arvot aktiviteetilla “Ex”. Tuloksena oli siis variantteja, kuten {A->B1->C->D1->Ex}. Mikäli tämä anonymisointitapa saataisiin kos-

kemaan vain harvinaisia variantteja, voisi se tarjota hyödyllisemmän anonymisoidun aineiston joitain käyttötapauksia varten. Tutkimuksessa käytetty DistributionGenerator-ohjelma ei tunnistanut Ex-tapahtumaa ja vaatisi siis tältä osin jatkokehitystä. Anonymisointi kohdistetaan vain harvinaisiin variantteihin ja DistributionGenerator-ohjelman jatkokehitys toteutetaan kokeessa 3.

6.1.3 Anonymisointi 3: Winsorizing-tyyppinen sensurointi

Lopulta kokeessa 1 päädyttiin käyttämään karkeaa winsorizing-tyyppistä sensurointia. Aineistosta siis poistettiin rivit, joiden polkuattribuutti oli yksilöivä k-attribuutilla 10. Tämä tarkoitti käytännössä kaikkia {A->B2->C->D2->E#} muotoisia variantteja. Todellisessa käytössä varianttien poisto vääristäisi tuloksia, joita saataisiin anonymisoitua dataa analysoimalla. Tuloksena sensuroinnista saatiin aikaleimaton anonymisoitu versio aineistosta, joka löytyy liitteestä A tunnisteella ”_timeless”.

Aikaleimaton anonymisoitu aineisto ja DistributionCollector-ohjelmalla kerätty jakaumadata syötettiin DistributionGenerator-ohjelmalle. Ohjelma lisäsi aineiston polkuattribuuttiin jakaumiin perustuvat kestot lisäten näihin Laplace-jakauman satunnaisuutta epsilon-arvolla 1. Tämän operaation tuloksena oli anonymisoitu polkuattribuutti-muodossa oleva aineisto, joka on saatavilla liitteestä A tunnisteella ”timeless_retimed”.

Anonymisoitu polkuattribuutti-muodossa oleva aineisto annettiin syötteenä TimestampNormalizer-ohjelmalle. Ohjelma tuotti anonymisoidun version aineistosta normaalissa tapahtumadata-muodossa, joka on saatavilla liitteestä A tunnisteella ”timeless_retimed_normalized”.

6.2 Tulokset

Generoitu lähtöaineisto saatiin polkuattribuutti-lähestymistapaa käyttämällä anonymisoitua suunnitellusti. Lähestymistapa kuitenkin rajoitti käytettäviä anonymisointimenetelmiä

merkittävästi. Osa aiheutuneista rajoituksista voi olla mahdollista kiertää soveltamalla aineiston esi- ja jälkikäsittelyä haluttuun anonymisointimenetelmään sopiviksi.

Automaattinen (k=2)-anonymisointi (right-to-left) ei onnistunut käytetyllä ohjelmalla halutusti. Anonymisointi kohdistui harvinaisten rivien sijasta kaikkiin riveihin. Koetta 2 varten luodaan oma anonymisointityökalu, joka kohdistaa vastaavan anonymisoinnin vain niihin riveihin, jotka eivät täytä k-ehtoa.

DistributionGenerator generoi uudet siirtymä- ja kestoajat myös niille riveille, joita anonymisointi ei muuttanut. Tämä on anonymisoinnin kannalta tarpeetonta ja muuttaa dataa turhaan. Kokeessa 2 tämä pyritään estämään muuttamalla DistributionCollector- ja DistributionGenerator-ohjelmien toimintaa, siten, että ne säilyttävät anonymisoinnista muuttumattomina selvinneiden rivien alkuperäiset siirtymä- ja kestoajat. Lisäksi muutetaan DistributionGenerator-ohjelmaa siten, että anonymisoinnissa muuttuneiden varianttien kestot generoidaan niitä mahdollisimman paljon vastaavien muuttumattomien varianttien pohjalta.

7 Koe 2: Löydettyjen ongelmien korjauksia

Koetta 2 varten muokattiin sekä DistributionCollector- että DistributionGenerator-ohjelmia kahdella tavalla. Yhtäältä siten, että anonymisoinnissa ennallaan pysyneet rivit säilytettiin ennallaan myös uusia kestoja generoitaessa; toisaalta siten, että anonymisoinnissa muuttuneiden rivien generoitavat kestot perustuvat vain niitä eniten muistuttaviin riveihin. Näiden lisäksi luotiin oma ohjelma automaattista (k=2) anonymisointia varten.

7.1 Oma anonymisointiohjelma

Lyhyen katsauksen jälkeen emme löytäneet ohjelmaa, joka pystyisi helppokäyttöisesti anonymisoimaan vain harvinaisia arvoja. Koska tälläinen toiminnallisuus on kohtuullisen yksinkertainen suorittaa *k*-anonymisoinnille, toteutimme ohjelman itse. Tällöin automatisoitu *k*-anonymisointi saadaan kohdistettua vain niihin riveihin, joilla on harvinaisia arvoja. Näin anonymisoidut tiedot pystyttiin palauttamaan tiedon normaaliin muotoon DistributionGenerator-ohjelman pienillä muokkauksilla. Tutkimuksessa käytetyssä datassa harvinaisia arvoja olivat polkuattribuutit, joiden kuvaama variantti oli muotoa {A->B2->C->D2->E#}.

7.2 Muokkaukset DistributionCollector- ja DistributionGenerator-ohjelmiin

7.2.1 Anonymisoinnissa muuttumattomien rivien säilyttäminen ennallaan

Menetelmää kehitettiin siten, että se säilyttää ennallaan niiden rivien aikavälit, joiden polkua anonymisointivaihe ei muuta. Tämä tehtiin säilyttämällä aikavälillinen polku datassa aikavälittömän ohella ja vertaamalla näitä anonymisointivaiheen jälkeen. Mikäli polun aktiviteetit ja niiden järjestys on sama, voidaan aikavälitön polkuattribuutti korvata lähdedatasta muuttumattomalla polkuattribuutilla.

Otetaan esimerkiksi polkuattribuuttimuodossa oleva rivi, joka esitetty taulukossa 16.

ID	StartTime	Path
0	2017-03-20 22:14:00	A(105):(2):B1(25):(61):C(64):(147):D1(64):(1020):E1(44)

Taulukko 16. Polkuattribuutissa oleva rivi.

DistributionCollector luo riville uuden TimelessPath-solun. Uuden solun arvo koostuu Path-solun sisältämistä aktiviteeteista ilman kestotietoja. Anonymisointi kohdistetaan TimelessPath-sarakkeeseen, kuten esitetty taulukossa 17.

ID	StartTime	Path	TimelessPath
0	2017-03-20 22:14:00	A(105):(2):B1(25):(61):C(64):(147):D1(64):(1020):E1(44)	A:B1:C:D1:E1

Taulukko 17. Polkuattribuutissa oleva rivi uudella kestotiedottomalla sarakkeella.

Kyseinen rivi ei muutu anonymisoinnissa, koska {A->B1->C->D1->E1} on yleinen variantti. DistributionGenerator vertaa anonymisoinnin jälkeen rivin Path- ja TimelessPath-arvoja. Mikäli kaikki aktiviteetit ovat samoja ja samassa järjestyksessä, DistributionGenerator käyttää alkuperäistä Path-arvoa sen sijaan, että se generoisi riville uusia kestoaikoja. Siirtymät jätetään huomiotta, koska kussakin aktiviteettien välissä oletetaan olevan siirtymä. Lopputulos on esitetty taulukossa 18.

Alkuperäinen polkuelementti	**Anonymisoinnin läpikäynyt polkuelementti**	**Täsmääkö**
A(105)	A	Kyllä
(2)	-	Ei
B1(25)	B1	Kyllä
(61)	-	Ei
C(64)	C	Kyllä
(147)	-	Ei
D1(64)	D1	Kyllä
(1020)	-	Ei
E1(44)	E1	Kyllä

Taulukko 18. Taulukon 17 polkuattribuuttien vertailu.

Koska rivi kulki anonymisoinnin läpi muuttumattomana, sen kuvaaman prosessi-instanssin rakenne ei muuttunut. Kaikki aktiviteetit ovat samoja ja samassa järjestyksessä, kuin ennen anonymisointia. Tästä seuraa, että rivi on DistributionGeneratorilla käsiteltynä identtinen suhteessa siihen, millainen se oli ennen DistributionCollectorilla käsittelyä, josta on esimerkki esitettynä taulukossa 19.

ID	**StartTime**	**Path**
4048	2018-01-12 03:20:00	A(54):(8):B2(60):(62):C(84):(81):D2(5):(39):E1(195)

Taulukko 19. Esimerkkirivi DistributionGenerator-ohjelman läpiajon jälkeen.

DistributionCollector lisää TimelessPath-arvon samoin, kuin edellisessä esimerkissä, joka on esitetty alla taulukossa 20.

ID	StartTime	Path	TimelessPath
4048	2018-01-12 03:20:00	A(54):(8):B2(60):(62):C(84):(81):D2(5):(39):E1(195)	A:B2:C:D2:E1

Taulukko 20. Esimerkkirivi DistributionGenerator-ohjelman läpiajon jälkeen.

{A->B2->C->D2>E1} on harvinainen variantti, joten se muuttuu anonymisoinnissa muotoon {A->B2->C->D2->Ex}, joka on esitetty alla taulukossa 21.

ID	StartTime	Path	TimelessPath
4048	2018-01-12 03:20:00	A(54):(8):B2(60):(62):C(84):(81):D2(5):(39):E1(195)	A:B2:C:D2:Ex

Taulukko 21. Esimerkkirivi DistributionGenerator-ohjelman läpiajon ja anonymisoinnin jälkeen.

DisributionGenerator tarkistaa ovatko aktiviteetit ja niiden järjestys säilyneet samana anonymisoinnin yli. Tarkistuksessa havaitaan, että viimeinen tapahtuma on muuttunut aktiviteetista E1 pseudoaktiviteetiksi Ex. Anonymisoidun hoitopulun vertailu alkuperäiseen on esitetty taulukossa 22.

Alkuperäinen polkuelementti	Anonymisoinnin läpikäynyt polkuelementti	Täsmääkö
A(54)	A	Kyllä
(8)	-	Ei

B2(60)	B2	Kyllä
(62)	-	Ei
C(84)	C	Kyllä
(81)	-	Ei
D2(5)	D2	Kyllä
(39)	-	Ei
E1(195)	Ex	Ei

Taulukko 22. Alkuperäisen ja anonymisoidun hoitopolun vertailu polkuelementeittäin.

DistributionGenerator kerää kestojakaumat varianttia {A->B2->C->D2->Ex} varten niistä varianteista, joiden rakenne vastaa sitä tarkimmin. Koska pseudoaktiviteettia Ex ei ole alkuperäisessä aineistossa, lähimmät variantit ovat kaikki muotoa {A->B2->C->D2->E#} olevat variantit. Tuloksena saatu rivi muistuttaa alkuperäistä riviä, mutta yksilöivä variantti {A->B2->C->D2->E1} on korvattu pseudovariantilla {A->B2->C->D2->Ex}, jonka aikaleimat on tuotettu jakaumasta. Käsittelyn tuloksena syntynyt rivi on kuvattu taulukossa 23.

ID	**StartTime**	**Path**
4048	2018-01-12 03:20:00	A(129):(5):B2(67):(241):C(175):(172):D2(76):(315):Ex(144)

Taulukko 23. Anonymisoitu esimerkkirivi.

7.3 Kokeen eteneminen

Kokeessa käsiteltävänä datana käytettiin samaa polkuattribuuttimuodossa olevaa pseudodataa kuin kokeessa 1. Data syötettiin DistributionCollector-ohjelmalle, joka tuotti jakaumadatatiedoston, sekä version datasta, joka sisälsi myös aikaleimattoman polkuattri-

buutti-sarakkeen. Jakaumadatatiedosto on identtinen kokeessa 1 tuotetun jakaumadatatiedoston kanssa.

Aikaleimattoman polkuattribuutin sisältävä versio datasta vietiin luodulle KAnonymizer-ohjelmalle, joka anonymisoi harvinaiset arvot aikaleimattomasta polkuattribuuttisarakkeesta.

Anonymisoitu data ja DistributionCollector-ohjelman luoma jakaumadata syötettiin DistributionGenerator-ohjelmalle. DistributionGenerator-ohjelma generoi uudet aikaleimat anonymisoinnin muuttamille riveille. Anonymisoinnissa muuttumattomat rivit säilyivät ennallaan.

7.4 Tulokset

Muokattu menetelmä toimii, mutta diskreettien empiiristen jakaumien käyttö tekee generoitujen arvojen hajonnasta suurta. Jakaumien sovitus (eng. distribution fitting) voisi onnistua, mutta on kohtuullisen monimutkaista ja aikaavievää. Kokeessa käytetyssä pseudoaineistossa, jakaumien sovituksen tilalta, saataisiin jakaumat muutettua suoraan eksponentiaali-jakaumaksi, koska kaikki ajat on alunperin generoitu eksponentiaali-jakaumasta. Todellisilla datoilla kestojen pohjalla olevat jakaumat voivat olla paljon monimutkaisempia (esimerkiksi Costa A., Jr (2017)), joten jakaumia ei tutkimuksessakaan tulkittu suoraan eksponentiaali-jakaumiksi. Esimerkki tuloksista on esitetty alla taulukoissa 24, 25 ja 26.

Alkuperäinen	**Anonymisoitu**
4048;2018-01-12 03:20:00;A(54):(8):B2(60):(62):C(84):(81):D2(5):(39):E1(195)	4048;2018-01-12 03:20:00;A(129):(5):B2(67):(241):C(175):(172):D2(76):(315):E#(144)
35331;2017-12-22 02:29:00;A(65):(13):B2(55):(130):C(93):(310):D2(89):(483):E3(16)	35331;2017-12-22 02:29:00;A(435):(26):B2(33):(39):C(2554):(373):D2(58):(315):E#(4)
57835;2017-08-18	57835;2017-08-18

14:00:00;A(8):(2):B2(20):(108):C(34):(634):D2(2):(244):E0(31)	14:00:00;A(212):(1):B2(150):(321):C(267):(366):D2(6):(791):E#(64)
77562;2017-02-06 09:50:00;A(10):(2):B2(14):(146):C(103):(44):D2(18):(506):E2(210)	77562;2017-02-06 09:50:00;A(41):(29):B2(65):(133):C(137):(74):D2(1):(342):E#(122)

Taulukko 24. Anonymisoimattomat ja anonymisoidut rivit polkuattribuuttimuodossa.

ID	**Act**	**StartTime**	**EndTime**
4048	A	2018-01-12 03:20:00	2018-01-12 04:14:00
4048	B2	2018-01-12 04:22:00	2018-01-12 05:22:00
4048	C	2018-01-12 06:24:00	2018-01-12 07:48:00
4048	D2	2018-01-12 09:09:00	2018-01-12 09:14:00
4048	E1	2018-01-12 09:53:00	2018-01-12 13:08:00
35331	A	2017-12-22 02:29:00	2017-12-22 03:34:00
35331	B2	2017-12-22 03:47:00	2017-12-22 04:42:00
35331	C	2017-12-22 06:52:00	2017-12-22 08:25:00
35331	D2	2017-12-22 13:35:00	2017-12-22 15:04:00
35331	E3	2017-12-22 23:07:00	2017-12-22 23:23:00
57835	A	2017-08-18 14:00:00	2017-08-18 14:08:00
57835	B2	2017-08-18 14:10:00	2017-08-18 14:30:00
57835	C	2017-08-18 16:18:00	2017-08-18 16:52:00
57835	D2	2017-08-19 03:26:00	2017-08-19 03:28:00
57835	E0	2017-08-19 07:32:00	2017-08-19 08:03:00
77562	A	2017-02-06 09:50:00	2017-02-06 10:00:00

77562	B2	2017-02-06 10:02:00	2017-02-06 10:16:00
77562	C	2017-02-06 12:42:00	2017-02-06 14:25:00
77562	D2	2017-02-06 15:09:00	2017-02-06 15:27:00
77562	E2	2017-02-06 23:53:00	2017-02-07 03:23:00

Taulukko 25. Anonymisoimattomat rivit normaalissa muodossa.

ID	**Act**	**StartTime**	**EndTime**
4048	A	2018-01-12 03:20:00	2018-01-12 05:29:00
4048	B2	2018-01-12 05:34:00	2018-01-12 06:41:00
4048	C	2018-01-12 10:42:00	2018-01-12 13:37:00
4048	D2	2018-01-12 16:29:00	2018-01-12 17:45:00
4048	E#	2018-01-12 23:00:00	2018-01-13 01:24:00
35331	A	2017-12-22 02:29:00	2017-12-22 09:44:00
35331	B2	2017-12-22 10:10:00	2017-12-22 10:43:00
35331	C	2017-12-22 11:22:00	2017-12-24 05:56:00
35331	D2	2017-12-24 12:09:00	2017-12-24 13:07:00
35331	E#	2017-12-24 18:22:00	2017-12-24 18:26:00
57835	A	2017-08-18 14:00:00	2017-08-18 17:32:00
57835	B2	2017-08-18 17:33:00	2017-08-18 20:03:00
57835	C	2017-08-19 01:24:00	2017-08-19 05:51:00
57835	D2	2017-08-19 11:57:00	2017-08-19 12:03:00
57835	E#	2017-08-20 01:14:00	2017-08-20 02:18:00
77562	A	2017-02-06 09:50:00	2017-02-06 10:31:00
77562	B2	2017-02-06 11:00:00	2017-02-06 12:05:00

77562	C	2017-02-06 14:18:00	2017-02-06 16:35:00
77562	D2	2017-02-06 17:49:00	2017-02-06 17:50:00
77562	E#	2017-02-06 23:32:00	2017-02-07 01:34:00

Taulukko 26. Anonymisoidut rivit normaalissa muodossa.

7.5 Generoitujen aikaleimojen jakaumat

Anonymisoituja rivejä löyty aineistosta neljä kappaletta. Anonymisoidut rivit ennen ja jälkeen anonymisoinnin ovat merkittynä taulukkoon 27.

A(54)	(8)	B2(60)	(62)	C(84)	(81)	D2(5)	(39)	E1(195)
A(206)	(59)	B2(65)	(166)	C(224)	(5)	D2(69)	(496)	E#(17)
A(65)	(13)	B2(55)	(130)	C(93)	(310)	D2(89)	(483)	E3(16)
A(104)	(19)	B2(151)	(215)	C(855)	(1952)	D2(20)	(613)	E#(302)
A(8)	(2)	B2(20)	(108)	C(34)	(634)	D2(2)	(244)	E0(31)
A(302)	(86)	B2(301)	(22)	C(747)	(625)	D2(29)	(1602)	E#(116)
A(10)	(2)	B2(14)	(146)	C(103)	(44)	D2(18)	(506)	E2(210)
A(63)	(36)	B2(73)	(287)	C(241)	(415)	D2(153)	(345)	E#(416)

Taulukko 27. Anonymisoidut rivit ennen ja jälkeen k=2 anonymisoinnin.

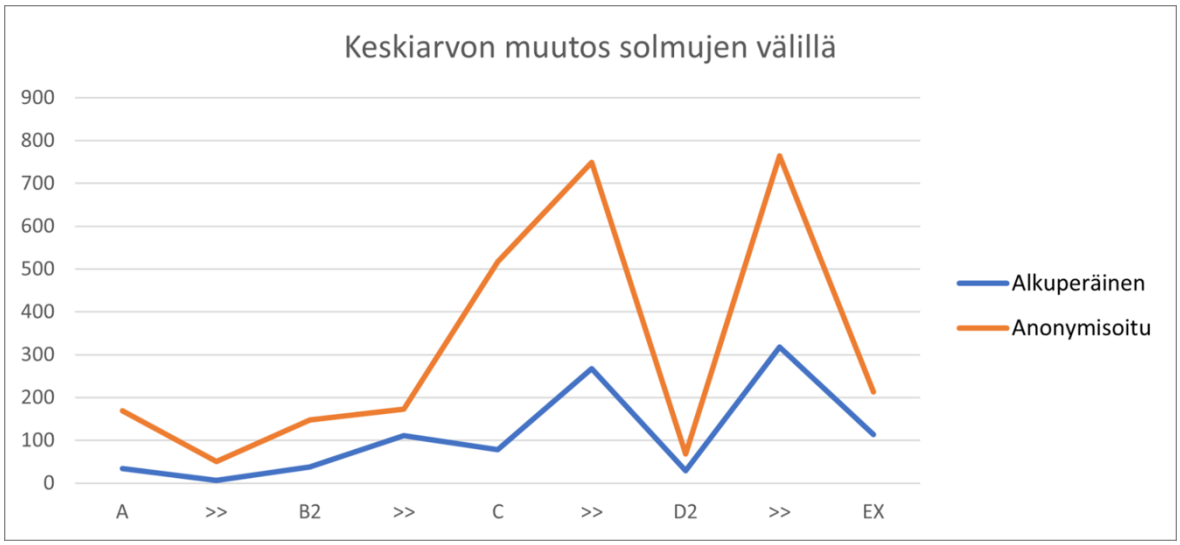
Anonymisoidut rivit muodostavat uuden pseudoryhmän, joiden kunkin aikaleimat otetaan huomioon aikavälien jakaumia luodessa. Tästä syystä aikaleimat voivat vaihdella suurestikin alkuperäisen ja anonymisoidun rivin välillä. Tätä vaihtelua todennäköisesti

voisi hillitä erinäisin menetelmin, mutta niiden toteuttaminen ei sisälly tähän tutkimukseen. Tilastollisia tunnuslukuja vaihtelusta on esitetty taulukossa 28.

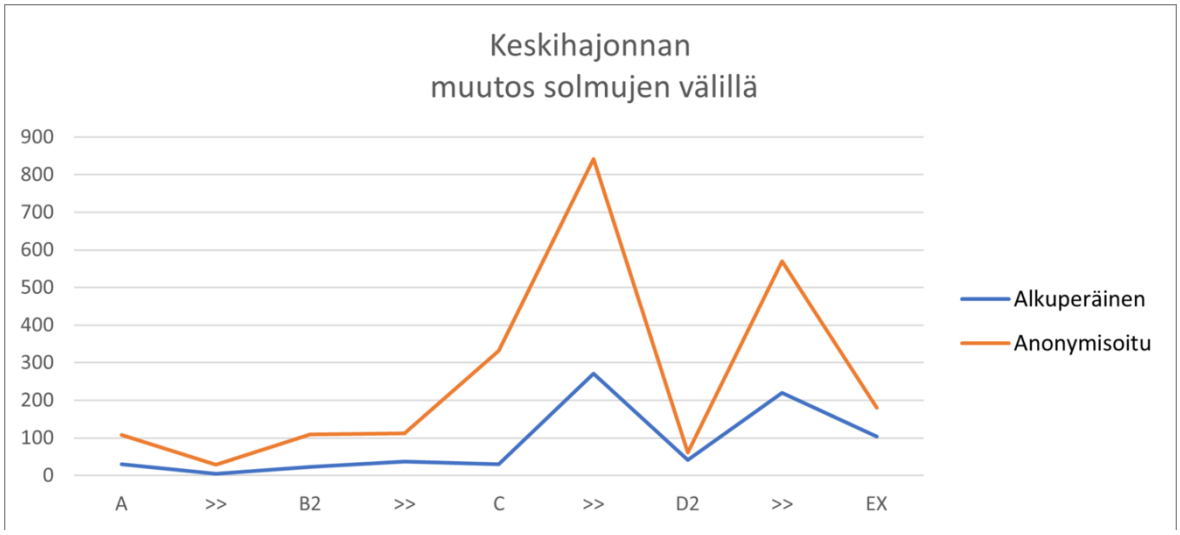
Alkuperäinen	**A**	**Väli**	**B2**	**Väli**	**C**	**Väli**	**D2**	**Väli**	**E**
Keskiarvo	34,25	6,25	37,25	111,50	78,50	267,25	28,50	318,0	113,0
Keskihajonta	29,51	5,32	23,60	36,50	30,66	271,33	40,93	220,52	103,71
Anonymisoitu	**A**	**Väli**	**B2**	**Väli**	**C**	**Väli**	**D2**	**Väli**	**E**
Keskiarvo	168,75	50,0	147,50	172,50	516,75	749,25	67,75	764,0	212,75
Keskihajonta	107,27	29,10	109,44	111,97	331,24	842,16	60,70	596,354	179,77
Erotus	**A**	**Väli**	**B2**	**Väli**	**C**	**Väli**	**D2**	**Väli**	**E**
Keskiarvo	-134,50	-43,75	-110,25	-61,0	-438,25	-482,0	-39,25	-446,0	-99,75
Keskihajonta	-77,76	-23,75	-85,84	-75,47	-300,58	-570,82	-19,77	-348,82	-76,06

Taulukko 28. Tilastollisia tunnuslukuja anonymisoiduille riveille ennen ja jälkeen anonymisoinnin.

Anonymisoidut arvot noudattelevat karkeasti alkuperäisten arvojen muutosta, joka esitetty kuvioissa 10 ja 11. Arvot korreloivat hyvin vahvasti kuten esitetty taulukon 29 arvoissa.



Kuvio 10. Keskiarvon muutos eri solmujen välillä alkuperäisen ja anonymisoidun aineiston välillä.



Kuvio 11. Keskihajonnan muutos eri solmujen välillä alkuperäisen ja anonymisoidun aineiston välillä.

Keskiarvon Pearson-korrelaatio	0,902161
Keskihajonnan Pearson-korrelaatio	0,93073

Taulukko 29. Keskiarvon ja -hajonnan korrelaatiot alkuperäisten ja anonymisoitujen rivien välillä.

Vaikka korrelaatio arvojen välillä on suurta, arvojen magnitudi vaihtelee suuresti. Tätä voitaisiin korjata erinäisin menetelmin, mutta ne eivät sisälly tutkielman laajuuteen. Korkea korrelaatio kuitenkin osoittaa datan soveltuvan hyvin tutkimuskäyttöön.

8 Pohdinta

Kokonaisuutena pro gradu -tutkielmamme keskiössä on sosiaali- ja terveydenhoidollisten tutkimusten kohteena olevien henkilöiden yksityisyyden suojan kehittäminen. Olemme tässä yhteydessä pohtineet erilaisia mahdollisia ja riittäviä ratkaisuja, joiden avulla voitaisiin huolehtia ja nostaa esille tutkimuksen kohteena olevien henkilöiden identiteettisuojaa koskevat seikat jo tutkimussuunnitelmia laadittaessa sekä tutkimusaineistojen data-analyysien suunnittelussa, tutkimusaineistojen käsittelyssä aina data-analyysin toteutukseen ja tulosten julkaisuun saakka. Yhtäaikaisesti on kyettävä huomioimaan useita eri tutkimuksellisia lähestymistapoja, ohjeistuksia, prosesseja ja säädöksiä, jotta tutkimuksen laadulliset ja eettiset kriteerit täyttyvät samalla kun kohderyhmän sisällä yksilön anonymiteetistä kyetään huolehtimaan myös suppeissa aineisto-otoksissa. Tutkimussuunnitelmat ovat lähtökohtana kaiken edellä kuvatun toteutumiseksi. Data-analyysin suunnittelu ja kuvaaminen tutkimussuunnitelmassa edellyttää tietoista pyrkimystä ja ymmärrystä jokaisen tutkimuksen kohteena olevan yksilön identiteetin suojaamiseen tutkimusaineiston käsittelystä julkistamiseen saakka. Itse asiassa sosiaali- ja terveysalan tutkimustoiminnassa ja sen suunnittelusta olisi hyvä noudattaa lääketieteen eettisenä ohjeena käytettävän Hippokrateen valan henkeä niin pitkälle kuin se on mahdollista. Eettisen ohjeistuksen lisäksi tutkimusaineiston keräämisessä ja käsittelyssä data-analyysiä varten on huomioitava voimassa oleva lainsäädäntö.

Sosiaali- ja terveysalalla kulloisenkin tutkimuksen kohderyhmä koko, toteutuksen alueellisuus, alueelliset rakenteet sekä palvelupolut asettavat omat anonymiteetin suojaamista edellyttävät toimenpiteet, erityisesti data-analyysin tulosten julkaisemisen jälkeen.

Olemme tässä pro gradu tutkielmassamme esittäneet muutamia mahdollisia tutkimuksellisia malleja miten sosiaali- ja terveysalan tutkimusten data-analyysi voitaisiin toteuttaa siten, että tutkimusten kohdehenkilöiden anonymiteetti olisi mahdollisimman hyvin suojattu koko tutkimusprosessin ajan - huomioiden sekä tutkimukselliset

laatukriteerit, eettiset ohjeistukset, voimassa oleva lainsäädäntö, erilaiset alueelliset rakenteelliset ratkaisut, palvelupolut.

8.1 Numeroituvien aikaleimojen käyttö

Monissa analyyseissä on tarpeellista tietää mille säännönmukaisesti toistuvalle ajanjaksolle tiettyjen aktiviteettien tapahtumat asettuvat. Esimerkiksi voidaan haluta tietää montako lonkkaleikkausta tehdään keskimäärin ensimmäisellä kvartaalilla tai minä viikonpäivänä esiintyy eniten tapaturma-diagnooseja. Tämänkaltaiset analyysit eivät tarvitse absoluuttisia aikaleimoja, vaan ne voidaan esittää joukkona numeroituvia (eng. enumerable) muuttujia. (Räsänen, Paavolainen, Sintonen, Koivisto, Blom, Ryynänen & Roine, 2009)

Voi siis olla mahdollista purkaa absoluuttiset aikaleimat valikoituun joukkoon numeroituvia muuttujia. Tämä mahdollistaisi *k*-anonymiteetin toteuttamisen myös aikaleimojen osalta. Numeroituvat muuttujat kannattanee valita tavalla, joka mahdollistaa halutut analyysit eivätkä ristiin verrattuina tuota absoluuttisia aikaleimoja. Esimerkiksi numeroituvat muuttujat kuun päivä, kuukausi ja vuosi[1] ilmaisevat suoraan absoluuttisen päivämäärän.

Sellaisen aineiston analysointi, jossa aikaleimat on esitetty numeerisina muuttujina voi vaatia erikoistuneiden analyysityökalujen luomista. On mahdollista, että monet käytössä olevista prosessianalyysi-työkaluista odottavat aikaleimojen esitystä absoluuttisina.

Polkuattribuutti-menetelmää käytettäessä tämä vaikuttaisi jakaumien keräämiseen ja uusien aikaleimojen generointiin. Sen sijaan, että kerättäisiin siirtymä- ja kestojakaumat, kerättäisiin kunkin aktiviteetin jakaumat kullekin halutulle numeroituvalle muuttujalle. Näiden pohjalta generoitaisiin kullekin anonymisoidulle pseudotapahtumalle omat numeroitavien muuttujien arvot.

[1] Vuosien arvo on kalenterissa kasvaa teoriassa äärettömästi, mutta käytännössä vuosia voidaan käsitellä numeroituvina.

8.2 Käytettyjen menetelmien ongelmia ja jatkokehityskohteita

Tutkimuksen aikana käytetyssä menetelmässä havaittiin useita ongelmia ja potentiaalisia jatkokehityskohteita. Osa näistä olisi ratkaistava ennen menetelmän käyttöä päätöksentekoa ohjaavissa analyyseissä.

DistributionGenerator-ohjelman koekäytössä havaittiin, että lähdeaineiston perusteella voi syntyä jakaumia, joiden keskihajonta on suuri suhteessa niiden keskiarvoon. Tällöin Laplace-jakauman mukaisen satunnaisuuden lisääminen jo valmiiksi arvaamattomaan arvoon voi johtaa tilanteisiin jossa aikavälitiedoksi generoituu negatiivinen arvo. DistributionGenerator-ohjelman arvon generointiin on tämän tilanteen varalta lisätty tarkistus, joka korvaa negatiiviset generoidut arvot uusilla. On kuitenkin huomattava, että arvojen korvaaminen aiheuttaa ainakin teoriassa vääristymää jakaumaan.

Toinen menetelmän ongelma on, ettei jakaumiin perustuvissa siirtymä- ja kestoajoissa huomioida rajoitteita, joita tiettyihin aktiviteetteihin voi kohdistua tosielämässä. Oletetaan esimerkiksi, että jokin terveysasema tekee kaikki laboratoriotutkimukset perjantaina. Koska jakauman pohjalta arvotaan laboratoriotutkimus-tapahtuman etäisyys sitä edeltävään tapahtumaan voi laboratoriotutkimus-tapahtuma päätyä esimerkiksi lauantaille. Tämänkaltaiset vääristymät voivat olla hyvin olennaisia joissain analyyseissä ja vaativat harkintaa jo ennen anonymisointia. Mahdollista ratkaisua tähän ongelmaan käsitellään edellisessä osiossa Numeroituvien aikaleimojen käyttö.

Lisäksi on huomattava, että tässä tutkimuksessa käytetty generoitu aineisto oli todelliseen aineistoon verrattuna hyvin yksipuolista. Tästä syystä differentiaalinen anonymisointi ei onnistunut. Menetelmän testaaminen monipuolisemmalla ja siten todellisenkaltaisemmalla aineistolla voisi tarjota mielenkiintoisia tuloksia.

9 Yhteenveto

Tutkielman tavoitteena oli löytää tai kehittää menetelmä, jolla terveydenhuollon hoitoprosessien yksilöiviä kulkuja saataisiin anonymisoitua tapahtumadatasta. Menetelmä saatiin kehitettyä ja sitä testattiin pseudodataa hyödyntämällä ensimmäisessä kokeessa. Menetelmää jatkokehitettiin ensimmäisessä kokeessa ilmenneiden puutteiden pohjalta. Tämän jälkeen menetelmää testattiin uudelleen.

Kehitetty polkuattribuutti-menetelmä on lupaava, mutta ainakin nykymuodossaan riittämätön monipuolisten palvelupolkujen hyödylliseen anonymisointiin. Parempia tuloksia voidaan saavuttaa kehittämällä menetelmää projektikohtaisesti.

Tutkimuksessa käytetty metodi voidaan tulkita riittäväksi ISO/IEC standardin 25237:2017 mukaisesti toteuttavan peruuttamattoman pseudonymisoinnin, mutta välttämättä saavuta ei EU:n GDPR:n tai ISO/IEC-standardin mukaista määritelmää anonymisoinnista, riippuen datan laadusta (ISO/IEC 25237:2017, GDPR). Toisaalta tämä menetelmä ei välttämättä ole tarpeellinen kaikkien maiden lainsäädäntöjen alla.

Anonymisaatio tarjoaa riittävän turvan yksityisyydelle, kun yksilöä ei voida aineistosta tunnistaa. Anonyymi henkilö tällaisessa tilastossa voi kuitenkin tulla tunnistetuksi, jos ensimmäisessä tilastossa anonymisoituja tietoja on yhdistettävissä toisen tilaston tietojen avulla anonymisoituun tilastoon. Tästä johtuen, jos julkaistaan useita aineistoja, jotka käsittelevät samoja henkilöitä on tärkeää budjetoida anonymisoinnin määrä siten, että anonymisaatio ei murru yhdistelemällä useita eri aineistojen tietoja. (Tang, Korolova, Bai, Wang & Wang, 2017)

Kirjallisuutta

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. The Journal of Law, Medicine & Ethics, 25(2-3), 98-110.

Kokkinakis, D. & Thurin, A. (2007). Anonymisation of Swedish Clinical Data. Conference on Artificial Intelligence in Medicine in Europe, AIME 2007.

Elger, B. S., Iavindrasana, J., Lo Iacono, L. Müller, H., Roduit, N., Summers, P., Wright, J. (2009). Strategies for health data exchange for secondary cross-institutional clinical research. Computer Methods and Programs in Biomedicine 99, 230-251

Alhazmi, A., & Arachchilage, N. A. G. (2021). I'm all ears! Listening to software developers on putting GDPR principles into software development practice. Personal and Ubiquitous Computing, 25(5), 879-892.

Mulder, T. (2019). Health apps, their privacy policies and the GDPR. European Journal of Law and Technology.

Tutkimuseettinen neuvottelukunta. (2013). Hyvä tieteellinen käytäntö ja sen loukkausepäilyjen käsitteleminen Suomessa. Saatavissa painettuna ja www-muodossa: https://tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf

Tutkimuseettinen neuvottelukunta. (2019). Ihmiseen kohdistuvan tutkimuksen eettiset periaatteet ja ihmistieteiden eettinen ennakkoarviointi Suomessa. Tutkimuseettisen neuvottelukunnan julkaisuja 3/19. Saatavissa painettuna ja www-muodossa: https://tenk.fi/sites/default/files/2021-01/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2020.pdf

Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9.

Ravindra, V., & Grama, A. (2021, June). De-anonymization attacks on neuroimaging datasets. In Proceedings of the 2021 International Conference on Management of Data (pp. 2394-2398).

Kalkman, S., van Delden, J., Banerjee, A., Tyl, B., Mostert, M., & van Thiel, G. (2022). Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. Journal of medical ethics, 48(1), 3-13.

Courbier, S., Dimond, R., & Bros-Facer, V. (2019). Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection-quantitative survey and recommendations. Orphanet journal of rare diseases, 14(1), 1-15.

Karampela, M., Ouhbi, S., & Isomursu, M. (2019, July). Exploring users' willingness to share their health and personal data under the prism of the new GDPR: implications in healthcare. In 2019 41st Annual International Conference of the Ieee Engineering in Medicine and Biology Society (embc)(pp. 6509-6512). IEEE.

ISO-standardi 25237:2017: Health informatics — Pseudonymization. Saatavissa www-muodossa: https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en

EU-asetus 679/2016: Yleinen tietosuoja-asetus. Saatavissa www-muodossa: https://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Kuula-Luumi, A. (2018). Turvaa tutkittavan anonymiteetti! Haettu 2022-04-09. Saatavissa www-muodossa: https://vastuullinentiede.fi/fi/jatkokaytto/turvaa-tutkittavan-anonymiteetti

Eronen, H. (2019). Käsitteletkö kuitenkin henkilötietoja? Haettu 2022-04-09. Saatavissa www-muodossa: https://vastuullinentiede.fi/fi/tutkimustyo/kasitteletko-kuitenkin-henkilotietoja

ISO/IEC-standardi 20889:2018: Privacy enhancing data de-identification terminology and classification of techniques. Saatavissa www-muodossa: https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en

Rocherm, L., Hendrickx, J. M., de Montjoye, Y-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications 10, artikkeli 3069.

Braunstein, S. L., Pati, A. K. (Julk. 2007). Quantum information cannot be completely hidden in correlations: implications for the black-hole information paradox. Physical Review Letters 98, 080502.

Wang, E. K., Jia, B., & Ke, N. (2017, December). Modeling background knowledge for privacy preserving medical data publishing. In 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC) (pp. 136-141). IEEE.

THE WORKING PARTY ON THE PROTECTION OF INDIVIDUALS WITH REGARD TO THE PROCESSING OF PERSONAL DATA. (2014). Opinion 05/2014 on Anonymisation Techniques. Saatavissa www-muodossa: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Podlesny, N. J., Kayem, A. V., Meinel, C., & Jungmann, S. (2019, November). How data anonymisation techniques influence disease triage in digital health: a study on base rate neglect. In Proceedings of the 9th International Conference on Digital Public Health (pp. 55-62).

Clarke, N., Vale, G., Reeves, E. P., Kirwan, M., Smith, D., Farrell, M., ... & McElvaney, N. G. (2019). GDPR: an impediment to research?. Irish Journal of Medical Science (1971-), 188(4), 1129-1135.

Negrouk, A., & Lacombe, D. (2018). Does GDPR harm or benefit research participants? An EORTC point of view. The Lancet Oncology, 19(10), 1278-1280.

Quinn, P. (2017). The Anonymisation of research data—A pyric victory for privacy that should not be pushed too hard by the EU data protection framework?. European Journal of Health Law, 24(4), 347-367.

Wierda, E., Eindhoven, D. C., Schalij, M. J., Borleffs, C. J. W., Amoroso, G., Van Veghel, D., ... & Ploem, M. C. (2018). Privacy of patient data in quality-of-care registries in cardiology and cardiothoracic surgery: the impact of the new general data protection regulation EU-law. European Heart Journal-Quality of Care and Clinical Outcomes, 4(4), 239-245.

Bayardo, R. & Agrawal, R. (2005). Data Privacy Through Optimal k-anonymization. IEEE: 21st International Conference on Data Engineering.

Aggarwal, C & Yu, P. (2008). Privacy-preserving Data Mining: Models and Algorithms. Kluwer Academic Publishers.

Antoniou, A., Dossena, G., MacMillan, J., Hamblin, S., Clifton, D., & Petrone, P. (2022). Assessing the risk of re-identification arising from an attack on anonymised data. arXiv preprint arXiv:2203.16921.

Ninghui, L, Tiancheng, L & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE: 23rd International Conference on Data Engineering.

Kairouz, P., Oh, S., Viswanath, P. (2014), Extremal Mechanisms for Local Differential Privacy. Advances in Neural Information Processing Systems 2014.

Sarathy, R. & Muralidhar K. (2011). Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. Transactions on Data Privacy 4, 1-17.

Domingo-Ferrer, J., & Soria-Comas, J. (2015). From t-closeness to differential privacy and vice versa in data anonymization. Knowledge-Based Systems, 74, 151-158.

Dwork, C. & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3-4.

Linsman, L, Rotter, T, James, E, Snow, P, Willis, J. (2010). What is a clinical pathway? Development of a definition to inform the debate. BMC Medicine 2010, 8: 31.

Pika, A., Wynn, M.T., Budiono, S., ter Hofstede, A.H.M., van der Aalst, W.M.P., Reijers, H.A. (2019). Towards Privacy-Preserving Process Mining in Healthcare. In: Di

Francescomarino, C., Dijkman, R., Zdun, U. (eds) Business Process Management Workshops. BPM 2019. Lecture Notes in Business Information Processing, vol 362.

Rafiei, M., von Waldthausen, L., van der Aalst, W. (2018). Ensuring Confidentiality in Process Mining.

Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M. (2019). PRETSA: Event Log Sanitization for Privacy-aware Process Discovery. International Conference on Process Mining (ICPM) 2019, 1-8.

Prasser, F., Eicher, J., Spengler, H., Bild, R., Kuhn, K. A., (2020). Flexible data anonymization using ARX - Current status and challenges ahead. Software: Practice and Experience, vol. 50, issue 7, 1277-1304.

Tukey, J. W., (1962), The Future of Data Analysis. The Annals of Mathematical Statistics, vol. 33, no. 1, 1-67.

Costa A., Jr (2017). Assessment of operative times of multiple surgical specialties in a public university hospital. Einstein (Sao Paulo, Brazil), 15(2), 200–205. https://doi.org/10.1590/S1679-45082017GS3902.

Räsänen, P., Paavolainen, P., Sintonen, H., Koivisto, A-M., Blom, M., Ryynänen, O-P., Roine, R. P., (2009), Effectiveness of hip or knee replacement surgery in terms of quality-adjusted life years and costs. Acta Orthopaedica vol. 78, 108-115.

Tang, J., Korolova, A., Bai, X., Wang, X., Wang, X., (2017), Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12. https://arxiv.org/pdf/1709.02753.pdf, haettu 28.07.2020.

Liitteet

A Tutkielmassa luodut ohjelmat

Saatavilla www-osoitteesta tai git-versionhallinta protokollalla
https://github.com/Tietun/progradu_path_attribute_anonymisation

B Luotujen ohjelmien dokumentaatiot

Tämä liite sisältää tutkielmaa varten luotujen ohjelmien dokumentaatiot. Kaikki ohjelmista on kirjoitettu Java-ohjelmointikielellä.

EHRDataGenerator

EHRDataGenerator on tutkielmassa käytetyn pseudo-datan luontiin käytetty ohjelma. Se sijaitsee datagenerator-paketissa, joka sisältää myös työkalu-luokat pseudo-asiakkaiden ja pseudo-palvelupolkujen luontiin. Ohjelma tulostaa pseudo-datan tiedostoon polkuattribuuttimuodossa.

EHRDataGenerator käyttää pääohjelmansa ohjelmakoodissa annettuja parametreja. Nämä on avattu seuraavassa taulukossa 30.

Parametrin nimi	**Parametrin tyyppi**	**Kuvaus**
ehrStart	LocalDateTime	A(5):(30):B(20)
ehrEnd	LocalDateTime	A(10):(30):B(20)
customerCount	int	A(5):(40):B(20)
carePathways	List<CarePathway>	Asiakkaille lisättävät palvelupolut
print	boolean	Tulostetaanko luotu data tiedostoon

filePath	String	Kohdetiedoston suhteellinen polku

Taulukko 30. EHRDataGenerator-ohjelman parametrit.

Generaattorin käyttämät CarePathway-oliot kuvaavat palvelupolkuja. Ne sisältävät kukin joukon järjestyksessä olevia Need-olioita, jotka kuvaavat kyseisen palvelupolun palvelutarpeita (EventNeed) ja palveluiden väliin jääviä aikavälejä (WaitNeed). CarePathway-oliolla on myös percentagePossibility-kenttä, joka määrittää kuinka todennäköisesti kyseinen palvelupolku annetaan asiakkaalle.

Generaattorin alkuperäinen tarkoitettu toiminta tapa oli luoda useita palvelupolkuja ja lisätä niitä asiakas-olioille (Customer) satunnaisuuteen perustuen. Tämä hylättiin kehityksen aikana, koska harvinaisia variantteja sisältävä yksittäinen palvelupolku oli helpommin hallittavissa ja hahmotettavissa. Muutoksen yhteydessä kyseinen palvelupolku kovakoodattiin CarePathway-luokan sisään.

EHRDataGenerator-ohjelman suoritus voi päättyä onnistuneesti tai teoriassa myös seuraaviin poikkeustilanteisiin, jotka on esitetty taulukossa 31.

Poikkeus	**Syy**
IOException	Tiedoston luku epäonnistui. Tämä voi johtua siitä, ettei määritettyä tiedostoa ole olemassa

Taulukko 31. EHRDataGenerator-ohjelman poikkeustilanteiden kuvaus.

DistributionCollector

DistributionCollector-ohjelma kerää Polkuattribuuttimuodossa olevasta datasta varianttikohtaiset kestojakaumat. Lisäksi se tuottaa datasta version, joka ei sisällä kestotietoja, sekä varinttikohtaiset kestotiedot. Kukin tuloste ohjataan omaan tiedostoonsa.

Ohjelma ottaa lähdetiedoston komentoriviparametrina. Tämän jälkeen ohjelma hyvin suoraviivaisesti kerää varianttikohtaiset tiedot, luo niiden pohjalta jakaumat ja tulostaa

nämä molemmat aikaleimattoman datan ohella tiedostoihin. Tulostetiedostojen nimet on asetettu ohjelmakoodissa ja syntyvät seuraavan taulukon 32 mukaisesti.

Tulostiedosto	**Nimi**
Aikaleimaton data	[lähdetiedoston nimi]_timeless.[lähdetiedoston tiedostopääte]
Variantit	[lähdetiedoston nimi]_variants.json
Kestojakaumat	[lähdetiedoston nimi]_distributions.json

Taulukko 32. DistributionCollector-ohjelman tulostiedostot.

DistributionCollector ohjelman suoritus voi päättyä onnistuneesti tai seuraavissa poikkeustilanteissa, jotka on esitetty taulukossa 33.

Poikkeus	**Syy**
IllegalArgumentException	Ohjeelle ei ole annettu tarvittuja tiedostopolku-argumentteja tai niiden polun muoto on väärä
FileNotFoundException	Jompaakumpaa tai kumpaakaan syötetiedostoa ei löytynyt
Muu IOException	Syötetiedostoja ei voitu lukea tai tulostiedostoon ei voitu kirjoittaa
Muu Exception	Virhe tiedostoja tulkitessa. Syötetiedostojen formaatissa voi olla vikaa. Vaihtoehtoisesti data-syötetiedosto sisältää yhä ainutlaatuisia palvelupolkuja.

Taulukko 33. DistributionCollector-ohjelman poikkeustilanteiden kuvaus.

DistributionGenerator

DistributionGenerator-ohjelma toimii vastakkaisesti suhteessa DistributionCollector-ohjelmaan. Se ottaa syötteenä aikaleimattoman polkuattribuutti-muodossa olevan datan ja DistibutionCollector-ohjelman luoman variantti-tiedoston. DistributionGenerator luo variantti-tiedostoon perustuvista jakaumista uudet aikaleimat syötedataan.

Uudet aikaleimat tuotetaan empiirisen diskreetin jakauman pohjalta. Jakaumasta saatuun arvoon lisätään Laplace-satunnaisuutta harvinaisten arvojen välttämiseksi. Laplace-jakauman lähtökohtainen epsilon-arvo on 1 (yksi). Epsilon-arvoa voidaan muuttaa DistributionGenerator-pääohjeman koodissa. DistributionGenerator-ohjelman virhetilanteet on avattu taulukossa 34.

Poikkeus	**Syy**
IllegalArgumentException	Ohjelle ei ole annettu tarvittuja tiedostopolku-argumentteja tai niiden polun muoto on väärä.
FileNotFoundException	Jompaa kumpaa tai kumpaakaan syöteetiedostoa ei löytynyt.
Muu IOException	Syötetiedostoja ei voitu lukea tai tulost tiedostoon ei voitu kirjoittaa
Muu Exception	Virhe tiedostoja tulkitessa. Syötetiedostojen formaatissa voi olla vikaa. Vaihtoehtoisesti data-syötetiedosto sisältää yhä ainutlaatuisia palvelupolkuja.

Taulukko 34. DistributionGenerator-ohjelman poikkeustilanteiden kuvaus.

TimestampNormalizer

TimestampNormalizer muuntaa polkuattribuutti-muotoisen datan normaaliin tapahtumaloki-muotoon.

KAnonymize

KAnonymize on yksinkertainen toteutus *k*-anonymisaatiosta, jota käytettiin kokeen kaksi *k*-anonymisoinnin toteuttamiseen. KAnonymize-ohjelman parametrit on kuvattu taulukossa 35.

Parametrin nimi	**Parametrin tyyppi**	**Kuvaus**
-f	String	Tiedostopolku *k*-anonymisoitavaan tiedostoon
-k	Int	*k*:n arvo
-c	String	Anonymisoitavan sarakkeen nimi

Taulukko 35. KAnonymize-ohjelman parametrit.

Ohjelma sensuroi valitun sarakkeen arvoja, joissa on vähemmän kuin k uniikkia arvoa oikealta vasemmalle käyttäen #-merkkiä, mutta ei muuta rivejä joista on yhtä monta tai useampi kuin k kappaletta identtisiä rivejä. Ohjelma tulostaa anonymisoidun sarakkeen arvot tiedostoon “out.txt” samaan kansioon, josta ohjelma ajettiin.