Synteettinen data terveysdatan yksityisyyshaasteiden ratkaisussa

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Kesäkuu 2026
Aurora Liukkonen

TURUN YLIOPISTO
Tietotekniikan laitos

AURORA LIUKKONEN: Synteettinen data terveysdatan yksityisyyshaasteiden ratkaisussa

LuK-tutkielma, 22 s.
Tietojenkäsittelytiede
Kesäkuu 2026

Tekoälymallit ovat edistäneet terveydenhuoltoa merkittävässä tahdissa, mutta niiden kehityksessä on haastavia esteitä. Terveysdatan jakaminen mallien kouluttamista varten on rajoitettua datan arkaluontoisuuden ja lainsäädännöllisten velvoitteiden takia. Synteettinen data on yksi tutkimuksissa esille noussut menetelmä datan yksityisyyden suojaamiselle.

Tutkielmassa tarkastellaan synteettisen datan käyttöä ja generointimenetelmiä terveydenhuollossa, keskittyen erityisesti terveysdatan yksityisyyshaasteisiin. Tutkielma toteutettiin kirjallisuuskatsauksena, jossa aineisto valittiin Web of Science ja ACM tietokantojen vuodesta 2020 eteenpäin julkaistuja tutkimuksia. Tutkielman tavoitteena on selvittää synteettisen datan tämänhetkiset hyödyt, haasteet ja mahdollisuudet lääketieteessä.

Tutkielmassa selviää synteettisen datan olevan monipuolisesti käytetty anonymisointimenetelmä, jolla on mahdollisuus helpottaa terveysdatan jakamista organisaatioiden välillä. Synteettisessä datassa on riski henkilötietojen uudelleentunnistuksesta ja yksityisyyttä vahvistavat menetelmät usein heikentävät synteettisen datan laatua huomattavasti. Synteettisen datan generointi- ja arviointimenetelmien jatkotutkimusta tarvitaan.

Asiasanat: synteettinen data, terveysdata, koneoppiminen, yksityisyys

Sisällys

Kuvat

Taulukot

1 Johdanto

Teknologian ja tekoälyn integroituminen osaksi terveydenhuoltoa on muuttamassa lääketiedettä datapainotteiseen ja digitalisoituun suuntaan. Tekoälyjärjestelmiä käytetään ja kehitetään laajasti terveydenhuollossa ja lääketieteellisessä tutkimuksessa tehokkaampaa ja yksilöidympää hoitoa varten. Tekoälymalleja hyödynnetään jokaisessa hoidon vaiheessa, esimerkiksi sairauksien ennustamisessa, kuvantamisessa ja resurssien hallinnassa. [1] Tekoälyn käyttö ja kehitys terveydenhuollossa edellyttää pääsyä alakohtaiseen dataan, joka ei ole helposti jaettavissa.

Terveysteknologian kehityksen ja yksilön yksityisyysoikeuksien välinen tasapainottelu on haastavaa. Digitalisoitumisen myötä datan määrä, mukaan lukien lääketieteellisen datan, on kasvanut eksponentiaalisesti. Terveysdatan avulla on mahdollisuus kehittää terveydenhuoltoa ja lääketieteellistä tutkimusta, mutta sen käyttö on rajattua. Terveysdata, kuten potilastiedot ja terveyshistoria, on arkaluontoista dataa, jonka käsittelyssä on velvollisuus noudattaa eettisiä ja lainsäädännöllisiä normeja. [2]

Datan perusteellinen ja tehokas suojaaminen ei ole yksinkertaista. Perinteiset suojakeinot, kuten tunnisteiden poistaminen, eivät välttämättä ole riittäviä ehkäisemään yksilön uudelleentunnistamista. Toisaalta liian suppea, vinoutunut tai karsittu koulutusdata heikentää sen monimuotoisuutta ja laatua, jolloin koneoppimismallien tarkkuus ja suorituskyky myös kärsii. Haasteena on datan tarkkuuden, hyödyllisyyden ja yksityisyyden tasapainottelu, mikä erityisesti tulee huomioida terveysdatas-

sa.[3] Turvallisuuden tehostaminen edellyttää oikeudenmukaisuuteen pohjautunutta tiedonhallintaa, kansainvälistä yhteistyötä ja yksityisyyteen pohjautuneiden järjestelmien kehittämistä [4]. Euroopan unionin eurooppalainen terveystietoalue (engl. European Health Data Space, EHDS) on esimerkki modernista viitekehyksestä, joka pyrkii avoimeen ja turvalliseen terveysdatan jakamiseen tutkimusta varten.[2]

Terveysdataan liittyvä haaste voitaisiin ratkaista esimerkiksi käyttämällä synteettistä dataa. Synteettinen data eli algoritmisesti tuotettu data jäljittelee aidon datan tilastollisia ominaisuuksia, jolloin se on verrannollinen alkuperäisen datan laatuun. Synteettinen data ei aidon datan mukaisesti sisällä suoraan tunnistettavia henkilötietoja, jolloin yksityisyydensuoja turvataan ja samalla saadaan helposti jaettavaa, laadukasta dataa. Synteettisellä dataa generoidaan useilla menetelmillä ja sitä voidaan tuottaa laajasti eri tarpeisiin, kuten mallin testaukseen tai aineiston täydentämiseen. Lääketieteessä tämä tarkoittaa esimerkiksi synteettisiä potilastietoja. Haasteet yksityisyysriskeistä ja datan laadusta eivät kuitenkaan katoa, minkä takia synteettisen datan tutkimus ja kehitys on tärkeää.[5]

Synteettisen datan laatu ja turvallisuus riippuu siitä, kuinka perusteellisesti se jäljittelee alkuperäistä dataa. Uudelleentunnistuksen ja yksityisyyshyökkäysten riskit kasvavat, kun malli oppii koulutusdatan rakenteen liian perusteellisesti ja saattaa generoida yksittäisiä aitoja datapisteitä synteettiseen aineistoon. Toisaalta liian heikosti aitoa dataa jäljittelevä synteettinen data-aineisto ei ole tarpeeksi todenmukaista ollakseen käyttökelpoista. Haaste on myös terveysdatan laadun määrittämisessä, jonka arviointi on haastavaa ilman alan asiantuntemusta.[6] Synteettisen datan generointiin on kehitetty yksityisyyttä edistäviä menetelmiä ja arviointikehyksiä datan laadun ja turvallisuuden tasapainoiselle vertailulle.

Tutkielma keskittyy terveysalan koneoppimismallien koulutuksen yksityisyysuhkiin ja tarkastelee mahdollisia synteettisen datan ratkaisuja ja haasteita. Synteettisen datan monialaisen käytön ja tutkimuksen takia tutkielma on rajattu lääke-

tieteeseen ja yksityisyysriskeihin rajatumman mutta kattavan aineiston käsittelyä varten. Tutkielmassa on kaksi tutkimuskysymystä, joiden näkökulmista ongelmaa tarkastellaan: TK1: Mitä yksityisyysriskejä liittyy tekoälyn käyttöön terveysdatan käsittelyssä? TK2: Miten synteettistä dataa voidaan hyödyntää terveysdatan yksityisyysriskeissä ja mitä haasteita siihen liittyy?

Työn tavoitteena on määrittää keskeiset terveysdatan yksityisyysuhat, määritellä synteettisen datan perusteet ja arvioida sen mahdollisuuksia ja rajoitteita yksityissuojan näkökulmasta.

Tutkielma on toteutettu kirjallisuuskatsauksena. Tiedonhaussa on hyödynnetty systemaattista aineistonhakua, jossa aineisto on haettu tarkennetuilla hakusanoilla eri tietokannoista. Rajauskriteereinä haussa käytettiin vain englanninkielisiä tutkimuksia ja tarkasteluun on otettu erityisesti vuodesta 2020 eteenpäin julkaistuja, tuoreempia tutkimuksia. Tiedonhakuprosessia on havainnollistettu taulukolla 1.1.

Taulukko 1.1: Aineiston tiedonhakuprosessi

Tietokanta	Hakusanat	Tuloksia yhteensä	Tarkasteltavat julkaisut	Valitut julkaisut
Web of Science	”synthetic data” AND ”privacy” AND “medicine” Rajaus: englanti	183	26	7
ACM	”synthetic data” AND ”privacy” AND “medicine” Rajaus: englanti ja abstrakti	20	5	1

Tutkielma alkaa taustatiedon käsittelyllä ja etenee pääaineiston tarkasteluun ja pohdintaan. Luvussa 2 määritellään tutkimukselle olennaiset käsitteet, tekoäly ja terveysdata, sekä esitellään terveysalaan painottuvia tekoälyn yksityisyysriskejä. Synteettisen datan määritelmään ja generointimenetelmiin perehdytään luvussa 3. Luvussa 4 tarkastellaan synteettisen datan käyttöä terveydenhuollossa ja siihen liit-

tyviä mahdollisuuksia ja haasteita. Luvussa 5 pohditaan synteettisen datan rajoja ja tulevaisuudenkuvaa.

2 Tekoäly ja terveysdata

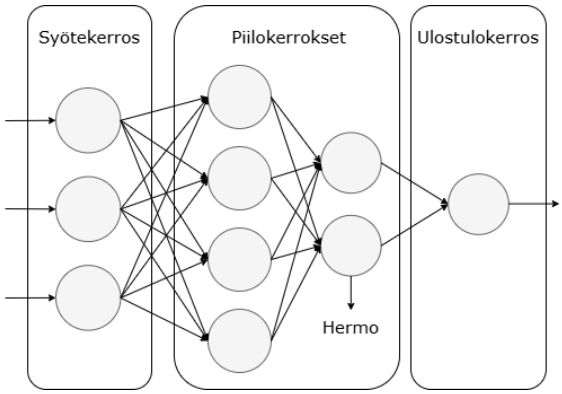
2.1 Tekoäly

Tekoälyllä ei ole yhtä yhteisesti hyväksyttyä määritelmää, sillä sana on muuttunut ajan kuluessa ja teknologian kehittyessä. Nykypäivänä tekoäly tuo mieleen varsinkin syväoppivat generatiiviset mallit, kuten ChatGPT ja Gemini. Toisaalta termi tarkoittaa myös yksinkertaisempia toimintoja suorittavia sääntöpohjaisia algoritmeja. Yleisesti tekoälyllä viitataan tietokonejärjestelmiin, jotka voivat itsenäisesti suorittaa ihmisälykkyyttä edellyttäviä tehtäviä [7]. EU:n tekoälysäädös (engl. AI Act) muotoilee tekoälyjärjestelmän tarkoittavan jollain autonomisuuden tasolla toimivaa tietokonejärjestelmää, joka syötteen pohjalta luo tuloksen [8].

Käsitteen laajuuden vuoksi tekoäly tyypillisesti jaetaan eri osa-alueisiin. *Koneoppiminen* on tekoälyn osa-alue, jossa tietokonealgoritmit pystyvät analysoimaan dataa ilman sääntöpohjaista ohjelmointia eli ilman ihmisen laatimia vastausvaihtoehtoja jokaisesta mahdollisesta tuloksesta. Nämä algoritmit eli koneoppimismallit koulutetaan suurilla datamäärillä tunnistamaan datan tilastollisia ominaisuuksia ja tekemään havaintoja, ennusteita tai päätöksiä sen perusteella. Koneoppimismenetelmät usein jaetaan ohjattuun, ohjaamattomaan ja vahvistettuun oppimiseen riippuen siitä, millä tavalla malli käsittelee koulutusdataa ja mikä sen tarkoitettu tehtävä on.[1]

Koneoppimismallien rakentaminen alkaa ratkaistavan ongelman määrittelyllä ja tarvittavien aineistojen etsinnällä, jotka siistitään ja muunnetaan koulutusdataksi soveltuvaan muotoon. Valittuja mallivaihtoehtoja koulutetaan aineistolla ja prosessi toistetaan, kunnes suorituskykyä on parannettu mahdollisemman paljon, jonka jälkeen parhaiten suoriutuva malli valitaan. Mikäli suorituskyky ei ole kelvollinen millään mallilla palataan aikaisempiin työvaiheisiin. [9]

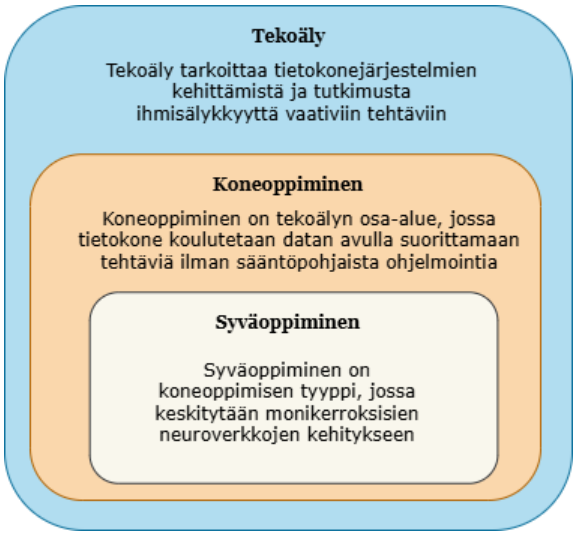
Syväoppiminen on toinen tekoälyn osa-alue, jossa algoritmeista rakennetaan ihmisaivojen toimintaa imitoivia monikerroksisia neuroverkkoja [1]. Neuroverkot koostuvat yksinkertaisimmillaan neuroneista eli keinotekoisista hermoista, jotka laskevat tuloksen ympäristöstä havaitulta tai toisilta neuroneilta saadusta ärsykkeestä [10]. Näin mallit oppivat datasta sen piirteitä ja osaavat soveltaa niitä tarvittuun tehtävään, kuten uuden datan luontiin. Esimerkki yksinkertaisesta perseptronineuroverkosta havainnollistetaan kuvassa 2.1.



Kuva 2.1: Monikerroksinen perseptroniverkko

Neuroverkkomallit ovat hyödyllisiä ratkaisukeinoja tilanteisiin, joissa perinteiset tekoälyalgoritmit vaatisivat monimutkaisia ja yksityiskohtaisia sääntöjä. Syväoppivat mallit ovat kuitenkin haasteellisia kouluttaa, sillä ne yleensä tarvitsevat suurempia koulutusaineistoja kuin perinteiset algoritmit. Varsinkin syväoppivien mallien erityishaaste on epäselvyys mallin tavasta päätyä lopputulokseensa ja tarve suurille,

vaikeasti saataville datamäärille.[11] Käsiteltyjä tekoälyn osa-alueita havainnollistetaan kuvassa 2.2.



Kuva 2.2: Tekoälyn osa-alueet. Muotoillen lähdettä [1]

Ilman dataa tekoälymalleja ei olisi mahdollista kouluttaa tai kehittää. *Koulutusdata* on tarkkaan valittu ja rajattu aineisto useasta eri lähteestä ja voi *datatyypiltään* olla esimerkiksi taulukoita, kuvia tai tekstidokumentteja. Runsas ja monimuotoinen aineisto edistää mallin tarkkuutta, luotettavuutta ja yksityisyyttä. Tarvittava datamäärä ja datan laatu riippuu myös vahvasti mallin käyttökohteesta ja vaaditusta yleistettävyydestä. [12]

2.2 Terveysdata

Tekoälyn hyödyntäminen terveydenhuollossa edellyttää suuria määriä alakohtaista dataa. Terveysdata tarkoittaa terveys- ja lääketieteenalaista tietoa yksilön terveydentilasta ja terveystiedoista, kuten hoidosta ja sairaushistoriasta. Terveysdataa ovat muun muassa digitaaliset potilaskertomukset (engl. electronic health records, EHR), geneettinen data, kuvantamis- ja laboratoriotulosdata. [13]

Terveysdatan käytössä tekoälymallien koulutukseen liittyy useita huomioonotettavia haasteita. Datan käsittelyyn tarvitaan alakohtaista ymmärrystä, sillä sen laadun arvioiminen ja vinoumien huomaaminen ovat haasteellisia terveysdatan kompleksisuuden takia. [14] Koulutusdatan ominaisuudet vaikuttavat tekoälymallin toimintaan, joten datan vinoumat ja kohina vaikuttavat suoraan mallin todenmukaisuuteen ja siihen, miten se erottaa erikoistapauksia. [12]

Toinen merkittävä haaste varsinkin terveysalan tekoälymallien kouluttamisessa on datan saatavuus. Koulutusdataksi tarkoitetun, julkisen datan määrä on niukkaa kysyntään verrattuna.[14] Tarvittava data koostuu useimmiten potilastiedoista, jotka ovat arkaluontoisia niiden henkilökohtaisen luonteen vuoksi. EU:n yleinen tietoturva-asetus GDPR määrittää kaiken yksilöstä kerättävän datan olevan tietosuojan alaista. GDPR tarkentaa terveyteen liittyvän datan olevan kaikkea henkilön fyysisestä ja psykologisesta terveydentilasta, geneettisestä ja biometrisestä informaatiosta kertovaa dataa.[15] Datan keräämisessä, käsittelyssä ja jakamisessa on siis moraalinen ja laillinen vastuu.

Terveysdatalla koulutetulla mallilla on yksilöintiriski potilaan uudelleentunnistuksesta, jossa ulkopuoliset tunnistavat yksilön hänen terveystietojensa perusteella, vaikka data olisi anonymisoitua.[16] Yksityisyyden vahvistamista varten on kehitetty menetelmiä, joita hyödynnetään koulutusdatan valmistelussa ja mallin koulutuksessa. Tilastolliset tietosuojamenetelmät (engl. statistical disclosure control, SDC) tarkoittavat teknisiä menetelmiä minimalisoida yksilön tunnistus datasta mahdollisemman perusteellisesti, jota kutsutaan myös datan anonymisoinniksi. Tämä tarkoittaa henkilöllisyyden paljastavien tunnisteiden, kuten nimen ja henkilötunnuksen, karsimista data-aineistosta. Kaikissa SDC menetelmissä tasapainotellaan aineiston hyödyllisyyden ja datan muokkaamisen välillä, sillä mallin tehokkuus ja luotettavuus kärsii, jos koulutusdata ei ole tarpeeksi runsasta ja todenmukaista. [17]

Tarve turvallisille ja datan laadun säilyttäville tietosuojamenetelmille on suuri etenkin terveydenhuollossa. Datan augmentaatio tarkoittaa tekniikoita kasvattaa data-aineiston kokoa ilman aidon datan keräämistä.[18] Siinä aitoa data-aineistoa täydennetään esimerkiksi muokkaamalla yksittäisiä datapisteitä ja lisäämällä ne alkuperäisen datan rinnalle. Yksinkertaisuudessaan se voi tarkoittaa kuvadatan editointia rajaamalla tai kääntämällä kuva.[12] Synteettinen data on yksi datan augmentaatiomenetelmä ja keino vahvistaa yksityisyyden suojaa, jota tarkastellaan luvuissa 3 ja 4.

3 Synteettinen data

Synteettinen data tarkoittaa keinotekoisesti luotua dataa, joka voi olla tilastollisen mallintamisen, generatiivisten mallien tai simulaatiomenetelmien avulla luotua. Tavoitteena on luoda dataa, joka on verrannollista aitoon dataan eli säilyttää sen tilastollisen rakenteen, jolloin dataa voidaan käyttää koneoppimisessa samoin kuin aitoa dataa. Synteettistä dataa käytetään muun muassa puutteellisen koulutusaineiston täydentämiseen ja koneoppimismallien testaukseen.[19] Yksityisyyshaasteiden näkökulmasta synteettinen data on varsin kiehtova, sillä generoitu data ei alkuperäisen datan tavoin ole aitoa henkilödataa. Tämä mahdollistaa data-aineiston jakamisen muiden käyttöön vapaammin kuin arkaluontoisen henkilödatan.[17]

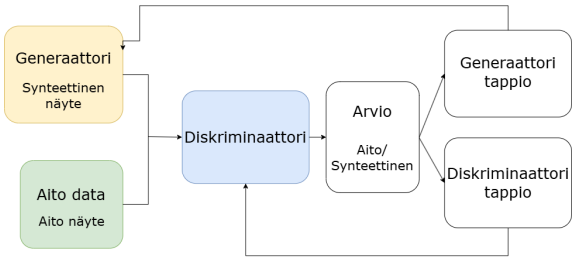
Synteettistä dataa on eri tasoista. *Täysin synteettinen data* tarkoittaa generoitua dataa, jonka tilastolliset piirteet ovat verrannollisia käytettyyn koulutusaineistoon. Edes täysin synteettinen data ei siis ole tyhjästä luotua, vaan mallin kouluttamiseen on käytetty aitoa dataa. *Osittain synteettinen data* on aitoa dataa, jonka arkaluontoisia tai puuttuvia osia on korvattu synteettisellä datalla esimerkiksi turvallisuussyistä.[17]

Riippumatta menetelmästä, synteettisen datan generointi alkaa data-aineiston keräämisestä. Data-aineistot koostuvat datapisteistä eli yksittäisistä datahavainnoista, mittausarvoista tai muusta tiedosta, esimerkiksi potilaan iästä tai sukupuolesta. Raaka eli valmistelematon data ei entuudestaan ole sopivassa muodossa mallin kouluttamista varten, joten aineisto täytyy muokata haluttuun muotoon. Valmistelulla

tarkoitetaan esimerkiksi eri aineistojen yhdistämistä suuremmaksi kokonaisuudeksi ja tyhjien datapisteiden poistamista. Koulutusdatan valmistuttua malli analysoi aineistoa tai se koulutetaan sillä. Viimeisessä vaiheessa mallin toimivuutta ja generoitua synteettistä dataa arvioidaan. [20]

Kuten edellä todettiin, synteettistä dataa voidaan generoida tilastollisilla, generatiivisilla ja simulaatiopohjaisilla menetelmillä. Tilastolliset menetelmät mallintavat datan todennäköisyysjakaumaa ja generoivat uutta dataa havaintojensa perusteella. Menetelmillä on etukäteen tietoa datan tilastollisista ominaisuuksista, toisin kuin syväoppivilla malleilla. Esimerkiksi *Bayes-verkko* esittää muuttujien riippuvuuksia, jolloin synteettiset otokset imitoivat alkuperäisen datan riippuvuuksia. Bayes-verkkoja voidaan kouluttaa eri ehdoilla, missä esimerkiksi verkko käsittelee koulutusaineistoa vain osittain. [21]

Generatiiviset kilpailevat neuroverkostot (engl. Generative Adversial Networks, GAN) ja *variationaaliset autoenkooderit (engl. Variational Autoencoder, VAE)* ovat generatiivisia malleja synteettisen datan luontiin. GANin arkkitehtuuri perustuu kahteen keskenään keskustelevaan neuroverkkoon, generaattoriin ja diskriminaattoriin. Generaattori generoi dataa ja diskriminaattorin tehtävä on arvioida onko data aitoa vai synteettistä. Diskriminaattori koulutetaan alustavasti aidolla koulutusdatalla tunnistamaan datan piirteitä, jota generaattori pyrkii generoimaan ilman koulutusdataa. Diskriminaattorin arvio antaa aina häviäjän ja voittajan; joko generaattorin generoima data ei ollut tarpeeksi todenmukainen tai diskriminaattori arvioi sen aidoksi. Neuroverkot toistavat prosessin useita kertoja ja muokkaavat parametrejaan saamansa palautteen mukaan, kunnes diskriminaattori ei kykene erottamaan aitoja datapisteitä synteettisistä. GAN-mallien heikkous on generaattorin pyrkimys johtaa diskriminaattori harhaan, jolloin datan monimuotoisuus kärsii ja malli generoi liian samankaltaisia datapisteitä. [21] Kuvassa 3.1 havainnollistetaan GAN-mallin toimintaa.



Kuva 3.1: GAN-malli

VAE on monikerroksinen perseptroniverkko eli autoenkooderi, joka koostuu enkooderista ja dekooderista. Enkooderi tiivistää datan latenttiavaruuteen ja dekooderi jälleenrakentaa pakatun datan, yrittäen jäljitellä alkuperäisen datan muotoa. VAE-mallien variaationaalisuus tarkoittaa latenttiavaruuden datapisteiden mallintamista todennäköisyysjakaumalla, joka mahdollistaa todennäköisyyslaskentaan perustuvan generatiivisuuden. [21]

Toisin kuin tilastolliset ja generatiiviset menetelmät, simulaatiopohjaiset menetelmillä mallinnetaan oikean maailman prosesseja ja mekanismeja eli yritetään generoida dataa miten se luonnollisesti syntyisi. Vahvistusoppimiseen tarvittavan ympäristön korvaaminen simulaatiolla on yleistä esimerkiksi robotiikassa, sillä se on huomattavasti halvempaa. Synteettistä dataa saadaan kerättyä simulaatiosta ja sitä myöhemmin voidaan tuunata aidolla datalla. [19]

Synteettinen data on houkutteleva erityisesti yksityisyyssuojan näkökulmasta, koska sitä voidaan jakaa ja käyttää vapaammin verrattuna alkuperäiseen terveysdataan. Tämä voi helpottaa ja nopeuttaa tekoälymallien kehittämistä merkittävästi, huomioimalla samalla eettiset rajoitteet ja yksilön oikeudet. [19]

Synteettiseen dataan liittyy haasteita ja riskejä, sillä sen laatu riippuu vahvasti alkuperäisestä datasta ja käytetyistä menetelmistä. Huonolaatuinen tai vinoutunut koulutusdata kykenee kouluttamaan vain käyttökelvottomia tekoälymalleja. Varsinkin synteettisessä datassa voi olla haasteellista arvioida ja todistaa sen todenmukaisuutta. Lisäksi synteettisen datan luonti on työlästä ja monimutkaista. Näiden

haasteiden takia synteettisen datan käyttö ei ole kovin laajaa, vaan sitä käytetään pääasiassa täydentämään puutteellista koulutusdataa tai testaamaan tekoälymalleja turvallisessa ympäristössä. Synteettisen datan käyttötarkoituksia terveydenhuollossa ja käytön mahdollisuuksia ja haasteita käsitellään tarkemmin luvussa 4.

4 Synteettisen datan käyttö terveysalalla

Luotettavien tekoälyjärjestelmien kehittäminen terveydenhuoltoa varten edellyttää suuria data-aineistoja, joiden saatavuus ja käsittely ovat haasteellisia terveysdatan eettisten rajoitteiden takia. Lainsäädännölliset ja organisaatiokohtaiset tietosuojasäädökset vaikeuttavat datan jakamista tutkimusyhteisöissä ja organisaatioiden välillä. Synteettinen data mahdollistaa terveysdatan jakamisen tilanteissa, joissa aito data ei ole saavutettavissa. [18], [22] Datan jakamisen lisäksi muita tutkimuksissa esille nousseita syitä synteettisen datan käytölle ovat mallien ennustavuuden vahvistaminen, mallin monipuolisen potilaskunnan huomioonoton tehostaminen ja resurssien säästäminen [20]. Tässä luvussa pyritään kartoittamaan synteettisen datan tämänhetkisiä käyttökohteita, mahdollisuuksia ja haasteita lääketieteellisessä tutkimuksessa ja terveydenhuollossa, keskittyen yksityisyyshaasteiden näkökulmaan.

Terveydenhuollossa synteettisellä datalla on useita käyttökohteita ja tutkimukset osoittavat laadukkaan synteettisen datan mahdollistavan analysoinnin ilman aitojen potilastietojen käsittelyä. Esimerkiksi GAN- ja VAE-mallit kykenevät luomaan todenmukaisia synteettisiä data-aineistoja riippumatta datatyypistä [5]. Synteettistä dataa voidaan hyödyntää harvinaisten ja aidoista aineistoista puuttuvien potilastapauksien simuloinnissa, joka on mallien tarkkuuden osalta tärkeää [20]. Sitä käytetään tekoälymallien kehittämisessä ja testauksessa, esimerkiksi korvaamalla arka-

luontoista kuvantamisdataa, kuten CT-skannaus- ja röntgenkuvia, tai EHR-dataa, kuten potilashistoriatietoja. [22].

Terveysdatan käsittely voi helpottua, mutta tutkimukset osoittavat synteettisessä datassa olevan huomioonotettavia yksityisyysriskejä. Vaikka synteettinen data ei suoraan kuvasta alkuperäistä dataa, yksilön uudelleentunnistus on mahdollista. Jos malli oppii koulutusdatan rakenteen liian tarkasti, se voi altistaa aitojen datapisteiden valumisen synteettiseen aineistoon. Mallit ovat riskialttiita paljastamiselle, jos koulutusdatana on ollut pieni tai epätasapainoinen aineisto, jolloin yksittäiset datapisteet ovat tunnistettavissa. Henkilötietojen epäsuora paljastuminen on erityisesti generatiivisten tekoälymallien haaste, joka tapahtuu mallin luodessa liian alkuperäistä koulutusdataa muistuttavaa synteettistä aineistoa. [23]

Suoran paljastamisen lisäksi mallit saattavat paljastaa yksilötietoja myös epäsuorasti. Synteettisen datan turvallisuus riippuu aineiston koon ja laadun lisäksi käytetyistä generointi- ja arviointimenetelmistä, sillä datasta on mahdollista päätellä alkuperäisen aineiston tietoa mallihyökkäysten avulla. Esimerkiksi membership inference -hyökkäys, jossa määritetään onko synteettinen datapiste periytynyt koulutusaineistosta. Hyökkäys kohdistuu koneoppimismalliin, mutta se myös kertoo synteettisen datan yksityisyyssuojan heikkoudesta. [5], [24]

Differentiaalinen tietosuoja (engl. differential privacy, DP) on matemaattinen viitekehys datan tilastollisten piirteiden julkaisuun samalla turvaten yksilön yksityisyyttä. DP algoritmit ovat suosittuja syväoppimismallien yksityisyyskoulutuksessa, mutta ne saattavat heikentää datan laatua. Useissa tutkimuksissa DP:n on huomioitu heikentävän synteettisen datan tilastollisia samankaltaisuuksia aitoon dataan merkittävästi [23]–[25]. Toisaalta DP:tä on myös käytetty ilman datan tarkkuuden menettämistä [5]. Esimerkiksi DP:llä koulutetut GAN-mallit suojaavat itseään membership inference -hyökkäykseltä tehokkaammin kuin ilman. [23], [25] Vaikka DP ei

aina ole suotuisin yksityisyyssuojatehostus mallin koulutuksessa, sen avulla voidaan myös saada tietoa alkuperäisestä aineistosta [23].

Mitä tarkemmin synteettinen aineisto jäljittelee alkuperäistä potilasdataa, sitä suurempi on mahdollinen riski alkuperäisten tietojen paljastumisesta. Samalla yksityisyyden vahvistaminen heikentää datan laatua, sillä se edellyttää datan yksityisyysriskialttiiden parametrien poistoa eli aineiston karsintaa. Tutkimuksissa tätä ongelmaa havainnollistetaan fidelity-privacy-utility-kehyksellä, jossa arvioidaan synteettisen datan todenmukaisuutta, yksityisyyssuojaa ja hyödyllisyyttä samanaikaisesti. Laadukasta synteettistä dataa tuottaessa tavoitellaan hyvää tasapainoa näiden painopisteiden välille, joka osoittautuu haastavaksi saavuttaa.[24] Tutkimuksissa synteettisen datan hyöty-yksityisyys-suhde on kuitenkin todettu suotuisemmaksi kuin anonymisoitu aito data [23].

Samojen yksityisyysongelmien pysyessä riippumatta synteettisen datan generointimenetelmistä herää kysymys voiko synteettistä dataa generoida ilman aitoa koulutusdataa. Synteettisen datan luominen ilman aitoa raakaa dataa on kuitenkin haasteellista ja epäkäytännöllistä [26]. Epätarkalla ja huonolaatuisella datalla on mahdotonta kouluttaa päteviä malleja, jolloin generoitu synteettinen data ei ole luotettavaa ja käyttökelpoista. Tämän takia synteettistä dataa itsessäänkään ei käytetä kouluttamaan synteettisen datan generointimalleja, sillä mallin suoriutumisissa on selvä ero aidolla data-aineistolla koulutettuun malliin.

Tutkimuksien perusteella synteettistä dataa sovelletaan monipuolisesti terveysalalla, varsinkin datan jakamisen, resurssien säästön ja yksityisyyssuojan edistämisen takia. Synteettinen data voi korvata aidon datan tietyissä tilanteissa, mutta tutkimuksissa havaitaan datan generoinnissa olevan samoja yksityisyysriskejä ja puutteita kuin mitä muissa anonymisaatiotekniikoissa esiintyy. Yksityisyyden vahvistaminen heikentää synteettisen datan laatua usein merkittävästi, joka esiintyy varsinkin DP:tä hyödyntävissä malleissa. Tutkimuksissa nousee esille tarve synteettisen da-

tan yksityisyysmenetelmien kehitykselle ja konkreettisille yksityisyyden ja hyödyn arviointiviitekehyksille olevan suuri.

5 Pohdinta

Tutkielma lähti mielenkiinnosta perehtyä dataan liittyviin yksityisyysongelmiin tarkemmin. Tällä hetkellä varsinkin EU:ssa keskustelu datasta ja yksityisyydestä on kiivasta, kun kamppailu yksityisyydensuojaa heikentävistä lainsäädöksistä, kuten Chat Control, ovat ajankohtaisimmillaan. Synteettinen data herätti mielenkiintoni tarkasteluaiheena, sillä se ensisilmäyksellä kuulosti utooppiselta teknologiaratkaisulta — liian hyvä ollakseen totta. Odotuksieni mukaan synteettinen data ei olekaan haasteeton ratkaisuvaihtoehto, vaan samat ongelmat kuin aidossakin datassa yhä pysyvät.

Synteettisen datan generointi ei poista alkuperäiseen dataan liittyviä yksityisyysriskejä. Yksityisyyshaasteet vain mukautuvat uuteen menetelmään, mahdollisesti jopa monimutkaistaen niitä. Esimerkiksi DP:n hyödyntäminen malleissa usein heikensi huomattavasti datan tarkkuutta, jolloin synteettinen data ei ollut käyttökelpoista huolimatta sen turvallisuudesta. Toisaalta eräissä tutkimuksissa DP ei vaikuttanut negatiivisesti datan laatuun ja paransi yksityisyyssuojaa, joten tutkimusten välillä on eroja. Tutkimuksissa käytettyjen menetelmien, mallien, aineistojen ja arviointimenetelmien eroavaisuudet vaikeuttivat niiden suoraviivaista vertailua keskenään. Ristiriitaisuudet viittaavat samoihin tarpeisiin, mitä tutkimuskirjallisuudessa yhtenevästi esiteltiin jatkotutkimustoiveina; synteettisen datan generointimenetelmien ja datan laadun arviointikehyksien kehittäminen.

Synteettisen datan tutkimus vaikuttaa yhä suhteellisen niukalta, vaikka useissa tutkimuksissa sen hyödyllisyys ja laatu todetaan lupaaviksi. Tutkimuksissa kehitetyt ja toivotut arviointikehykset keskittyivät kuitenkin samojen tavoitteiden ympärille eli tarkkuuden, hyödyllisyyden ja yksityisyyden ympärille. Tämä mahdollisesti osoittaa tutkimuksien olevan lähempänä yhteisen, pätevän arviointimenetelmän kehittämisessä kuin ei.

Onko mahdollista saavuttaa, tällaisen arviointiviitekehyksen mukaista, täydellisen synteettisen datan generointia tai onko se jotakin, mitä kuuluisi tavoitella? Tietysti jokaista menetelmää ja data-aineistoa on mahdollista hienosäätää ja muokata aivan atomitasolle asti, mutta onko tavoite aikaansaava ja realistinen. Riski virheestä voidaan, ja on hyväkin, minimalisoida, mutta ei poistaa. Jos tekoälyä halutaan hyödyntää terveysteknologiassa, vaara terveysdatan yksityisyysuhista tulee aina olemaan läsnä.

Pelkästään yhdessä tutkimuksessa tarkasteltiin synteettisen datan jatkokäyttöä ja evoluutiota. Synteettistä dataa ei voida, ainakaan tällä hetkellä, hyödyntää koulutusaineistona datan generointimallien koulutuksessa. Laadukasta uutta dataa ei pystytä luomaan loputtomasti mikäli pääsyä aitoon dataan ei ole. Tämä olisi mielestäni tärkeä ottaa huomioon jatkotutkimuksessa, sillä synteettisten aineistojen käyttö koulutusdatana voisi vahventaa yksityisyyttä entistä tehokkaammin.

Työn terveydenhuoltoon rajattu näkökulma ja valittu aineisto rajaavat myös työn näkemystä synteettisestä datasta ja sen haasteista, sillä siinä ei perehdytä tutkimukseen ja generointiin muussa kirjallisuudessa. Terveysdatan erityinen arkaluontoisuus vaikeuttaa tutkimusta ja tekoälyn kouluttamista enemmän kuin useilla muilla aloilla, jolloin synteettisen datan riskeissä ollaan myös kriittisempiä ja vaativampia. Aineistossa olisi täsmällisyyden edistämiseksi ollut hyvä valita myös vanhempia alan tutkimuksia, sillä se olisi tarjonnut paremman ymmärryksen synteettisen datan kehityksen saavutuksiin.

Tämän kirjallisuuskatsauksen perusteella voidaan todeta synteettisen datan olevan tehokas yksityisyysuhkia lieventävä keino terveydenhuollossa. Synteettistä dataa voidaan monipuolisesti generoida useilla eri menetelmillä ja yksityisyyttä tehostavilla menetelmillä. Synteettisen datan generoinnin erityisinä haasteina on tutkimuksen suppeus ja puute yleispätevästä arviointikehyksestä. Arvioinnissa tavoitellaan yksityisyyden, tarkkuuden ja hyödyllisyyden tasapainottamista, joka on haasteellista saavuttaa kaikissa anonymisaatiomenetelmissä. Synteettisen datan tutkimus on osoittanut lupaavia tuloksia useissa eri lääketieteen käyttökohteissa, mutta sen laadun arviointi ja kehitys tarvitsee jatkotutkimusta.

6 Yhteenveto

Tässä tutkielmassa tarkasteltiin synteettisen datan käyttökohteita terveydenhuollossa ja sen vaikutusta terveysdatan yksityisyyshaasteisiin. Aihetta käsiteltiin kahden tutkimuskysymyksen näkökulmasta: TK1: Mitä yksityisyysriskejä liittyy tekoälyn käyttöön terveysdatan käsittelyssä? ja TK2: Miten synteettistä dataa voidaan hyödyntää terveysdatan yksityisyysriskeissä ja mitä haasteita siihen liittyy?

Terveydenhuollon tekoälymallien kehittäminen on haastavaa terveysdatan arkaluontoisten henkilötietojen ja lainsäädännöllisten rajoitteiden takia. Datassa ja malleissa on riski uudelleentunnistaa yksilö tai käytetty aineisto, eivätkä anonymisaatiomenetelmät kykene turvaamaan yksityisyyttä täysin. Datan jakaminen organisaatioiden ja tutkimusyhteisöiden välillä on rajoitettua, joten tarve datan perusteellisen yksityisyyssuojan takaamiselle on suuri.

Tutkielmassa selviää synteettisen datan olevan lupaava yksityisyyssuojamenetelmä terveysdatan käsittelyyn, joka voi helpottaa datan jakamisen lisäksi datan monimuotoisuutta ja säästää resursseja. Yksityisyyshaasteet eivät kuitenkaan katoa synteettisen datan käytössä. Synteettisesti generoitu data saattaa paljastaa tietoja alkuperäisestä koulutusaineistosta, kasvattaen uudelleentunnistuksen riskiä. Synteettinen data on myös herkkä koulutusdatan laadulle, monimuotoisuudelle ja koolle, jolloin dataa karsivat yksityisyysvahvistukset heikentävät generoidun datan laatua merkittävästi. Keskeisimmät jatkotutkimustoiveet synteettisen datan yksityisyydes-

sä ovat tarve generointimenetelmien yksityisyyden ja hyöty-yksityisyys-suhteen arviointimenetelmien kehittämiselle.

Tämä kirjallisuuskatsaus oli pintaraapaisu synteettisen datan tutkimuksesta ja käytöstä terveydenhuollossa. Synteettinen data voi tutkitusti olla laadultaan verrannollinen aitoon dataan ja myös heikentää yksityisyysriskejä, mutta haasteena on tasapainottelu näiden välillä. Tutkimukset painottavat tarvetta synteettisen datan arviointikehyksille ja generointimenetelmien kehitykselle. Tutkielma korostaa tarvetta synteettisen datan jatkotutkimukselle ja painottaa terveysdatan yksityisyyssuojan tärkeyttä.

Lähdeluettelo

[1] M. Faiyazuddin et al., "The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of Advancements in Diagnostics, Treatment, and Operational Efficiency", *Health Science Reports*, vol. 8, nro 1, 2025. DOI: https://doi.org/10.1002/hsr2.70312.

[2] Euroopan unionin neuvosto, *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*, 2025.

[3] S. Mooney ja V. Pejaver, "Big Data in Public Health: Terminology, Machine Learning, and Privacy", *Annual Review of Public Health*, vol. 39, 2018. DOI: 10.1146/annurev-publhealth-040617-014208.

[4] A. Conduah, S. Ofoe ja D. Siaw-Marfo, "Data privacy in healthcare: Global challenges and solutions", *DIGITAL HEALTH*, vol. 11, 2025. DOI: 10.1177/20552076251343959.

[5] L. Yintong, A. Rajendra ja T. Jen Hong, "Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation", *Computer Methods and Programs in Biomedicine*, vol. 260, 2025. DOI: 10.1016/j.cmpb.2024.108571.

[6] E. Karoulla, N. Matragkas, S. U. Islam ja G. Epiphaniou, "Challenges and Risks of Using Synthetic Data for AI-Driven Healthcare Applications", *Stu-*

dies in Health Technology and Informatics, vol. 328, 2025. DOI: 10.3233/SHTI250677.

[7] S. Russell, *Artificial intelligence: a modern approach, Third edition, Global edition*. Pearson Education Limited, 2016.

[8] Euroopan unionin neuvosto, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*, 2024.

[9] S. Studer et al., "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology", *CoRR*, vol. abs/2003.05155, 2020. DOI: https://doi.org/10.48550/arXiv.2003.05155.

[10] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview", *Neural Networks*, vol. 61, 2015. DOI: 10.1016/j.neunet.2014.09.003.

[11] I. D. Mienye ja T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications", *Information*, vol. 15, 2024. DOI: 10.3390/info15120755.

[12] N. Nivedhaa, "A COMPREHENSIVE REVIEW OF AI'S DEPENDENCE ON DATA", *International Journal of Artificial Intelligence and Data Science*, vol. 1, 2024. DOI: 10.13140/RG.2.2.27033.63840.

[13] Euroopan unionin neuvosto, *Euroopan parlamentin ja neuvoston asetus (EY) N:o 1338/2008, annettu 16 päivänä joulukuuta 2008, kansanterveyttä sekä työterveyttä ja työturvallisuutta koskevista yhteisön tilastoista*, 2008.

[14] V. A. Rudrapatna ja A. J. Butte, "Opportunities and challenges in using real-world data for health care", *Journal of Clinical Investigation*, vol. 130, 2020. DOI: 10.1172/JCI129197.

[15] Euroopan unionin neuvosto, *Euroopan parlamentin ja neuvoston asetus (EU) 2016/679, annettu 27 päivänä huhtikuuta 2016, luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta (yleinen tietosuoja-asetus)*, 2008.

[16] T. Pham, "Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use", *Royal Society Open Science*, vol. 12, 2025. DOI: 10.1098/rsos.241873.

[17] T. E. Raghunathan, "Synthetic Data", *Annual Review of Statistics and Its Application*, vol. 8, 2021. DOI: 10.1146/annurev-statistics-040720-031848.

[18] J.-F. Rajotte, R. Bergen, D. L. Buckeridge, K. E. Emam, R. Ng ja E. Strome, "Synthetic data as an enabler for machine learning applications in medicine", *Iscience*, vol. 25, 2022. DOI: 10.1016/j.isci.2022.105331.

[19] S. I. Nikolenko, "Synthetic Data for Deep Learning", *CoRR*, 2019. DOI: 10.48550/arXiv.1909.11512.

[20] V. C. Pezoulas et al., "Synthetic data generation methods in healthcare: A review on open-source tools and methods", *Computational and Structural Biotechnology Journal*, vol. 23, 2024. DOI: 10.1016/j.csbj.2024.07.005.

[21] A. Bauer et al., "Comprehensive Exploration of Synthetic Data Generation: A Survey", *arXiv*, 2024. DOI: 10.48550/arXiv.2401.02524.

[22] X. Xing et al., "Non-imaging Medical Data Synthesis for Trustworthy AI", *Association for Computing Machinery*, vol. 56, 2024. DOI: 10.1145/3614425.

[23] T. Adams et al., "On the fidelity versus privacy and utility trade-off of synthetic patient data", *Iscience*, vol. 28, 2025. DOI: 10.1016/j.isci.2025.112382.

[24] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde ja N. Aginako, "Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees", *Frontiers in Digital Health*, vol. 7, 2025. DOI: 10.3389/fdgth.2025.1576290.

[25] V. Cheng, V. M. Suriyakumar, N. Dullerud, S. Joshi ja M. Ghassemi, "Can You Fake It Until You Make It?: Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness", teoksessa *Proceedings of the 2021 Acm Conference on Fairness, Accountability, and Transparency, Facct 2021*, Association for Computing Machinery, 2021. DOI: 10.1145/3442188.3445879.

[26] M.-A. Tammisto, F. A. Shah, D. Rodriguez ja D. Pfahl, "The Challenge of Generating and Evolving Real-Life Like Synthetic Test Data Without Accessing Real-World Raw Data-A Systematic Review", *Expert Systems*, vol. 42, 2025. DOI: 10.1111/exsy.70164.