Annika Kuoppamaa

Luotettavan tekoälyn haasteet ja ratkaisut terveydenhuollossa

Kirjallisuuskatsaus nykytilasta

Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tietojärjestelmätieteen kandidaattitutkielma
Kauppatieteiden kandidaatti

Vaasa 2025

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tekijä: Annika Kuoppamaa
Tutkielman nimi: Luotettavan tekoälyn haasteet ja ratkaisut terveydenhuollossa : Kirjallisuuskatsaus nykytilasta
Tutkinto: Kauppatieteiden kandidaatti
Oppiaine: Tietojärjestelmätiede
Työn ohjaaja: Teemu Mäenpää
Valmistumisvuosi: 2025 **Sivumäärä:** 49

TIIVISTELMÄ:

Tutkielmassa käsitellään luotettavaa tekoälyä terveydenhuollossa, mitä luotettavuus tässä kontekstissa tarkoittaa, mitkä tekijät siihen vaikuttavat, millainen tilanne on tällä hetkellä ja millaisia ratkaisuja luotettavuuden parantamiseksi on löydetty. Tekoälyn luotettavuutta arvioitiin Euroopan komission luomien eettisten periaatteiden pohjalta. Aiheen merkityksellisyyttä korostavat väestön ikääntymisestä johtuvat terveydenhuollon haasteet, globaali resurssipula terveydenhuollossa sekä tekoälyn kasvava rooli palveluiden laadun ja tehokkuuden parantamisessa. Työssä esille nousevat myös tekoälyn hyödyntämiseen liittyvät eettiset, tekniset ja yhteiskunnalliset kysymykset.

Tutkimus suoritettiin narratiivisena kirjallisuuskatsauksena. Tutkimusaineisto koostui viidestätoista vuosina 2021–2024 julkaistusta, Web of Science -tietokannasta valitusta tieteellisestä artikkelista. Artikkelit valittiin rajatuilla hakusanoilla ja seulottiin niiden relevanssin perusteella. Aineisto analysoitiin luokittelemalla ja teemoittelemalla. Työn tutkimuskysymykset tarkentuivat aineiston analyysin aikana ollen lopulta muotoa: Millaisia haasteita nykytilanteessa on havainnoitu luotettavan tekoälyn suhteen terveydenhuollossa? Millaisia ratkaisuja luotettavuuden parantamiseksi on löydetty?

Tutkimuksessa nousi esiin terveydenhuollon tekoälyn luotettavuudessa monia haasteita, kuten datan laadun ongelmia, algoritmien läpinäkymättömyyttä (mustan laatikon ongelma), organisaatiotason vastuiden laiminlyöntiä sekä eettisten periaatteiden ja käytännön soveltamisen välistä kuilua. Kaupalliset intressit havaittiin erityisen ongelmallisiksi, sillä ne vaikeuttavat läpinäkyvyyttä ja estävät tietojen kattavaa julkista raportointia. Tutkituissa artikkeleissa esitettiin ratkaisuiksi muun muassa moniammatillisen yhteistyön lisäämistä, tekoälymallien varhaisia ja jatkuvia auditointeja sekä erilaisia arviointimenetelmien kuten esimerkiksi Z-Inspection, MCAE-malli ja LIME- sekä Grad-CAM-tekniikoiden hyödyntämistä läpinäkyvyyden parantamiseksi. Tutkimuksessa korostuu tarve selkeille sääntelyvaatimuksille, läpinäkyville ja kattaville auditointiprosesseille sekä sidosryhmien väliselle avoimelle vuorovaikutukselle ja yhteistyölle. Tekoälyjärjestelmien suunnitteluun ja kehittämiseen tulee osallistaa moniammatillisia tiimejä jo varhaisessa vaiheessa. Lisäksi arviointikehyksiä tulee mukauttaa tarkasti käyttötilanteiden mukaan, ja tekoälyn eettisyyden ja luotettavuuden jatkuva empiirinen tutkimus on välttämätöntä teknologian laajamittaisen hyväksyttävyyden ja käyttöönoton varmistamiseksi terveydenhuollossa.

AVAINSANAT: tekoäly, terveydenhuolto, lääketiede, luottamus, luotettavuus, luotettava tekoäly

Sisällys

Kuviot

Taulukot

Lyhenteet

ALTAI: Assessment List for Trustworthy Artificial Intelligence. Euroopan komission perustaman korkean tason asiantuntijaryhmän (AI HLEG) luotettavaa tekoälyä koskevat eettiset ohjeet.

1 Johdanto

Tässä tutkimuksessa käsitellään luotettavaa tekoälyä erityisesti terveydenhuollon näkökulmasta. Tarkoituksena on perehdyttää lukijaa terveydenhuollossa käytettävän tekoälyn luotettavuuteen liittyviin teemoihin. Mitä luotettava tekoäly terveydenhuollossa tarkoittaa, mitkä tekijät siihen vaikuttavat, mikä on tilanne tällä hetkellä ja millaisia ratkaisuja luotettavuuden parantamiseksi on pyritty löytämään.

Aihealue on merkittävä niin kotimaassa kuin globaalisti. Sitra (Dufva, 2024) mainitsee vuoden 2024 megatrendeiksi hyvinvoinnin haasteiden kasvamisen ja kilpailun digivallasta. Teknologian kehitys on nopeaa, ja sitä otetaan käyttöön yhä laajemmin elämän eri osa-alueilla. Sitran mukaan tekoälyyn kiteytyy monta laajempaa kysymystä: miten dataa saa kerätä ja käyttää, miten käy yksityisyydelle ja tekijänoikeuksille, kenellä on paras osaaminen, resurssit, algoritmit ja patentit, sekä kuinka paljon nämä kehitettävät palvelut kuormittavat ympäristöä. Tekoälyn merkityksellisyyden rinnalle Sitra nostaa myös muun muassa kyberturvallisuuden sekä resilienssin eli kyvyn toimia muuttuvissa olosuhteissa, ja kohdata häiriöitä ja kriisejä. Bærøe ja muiden (2020, s. 260) mukaan tekoälyteollisuutta ohjaavat vahvat taloudelliset ja poliittiset intressit, ja siksi tekoälyn luotettavuuden tarve terveydenhuollossa on ratkaisevan tärkeää.

Suomessa hyvinvointia haastavat väestön ikääntyminen ja kasvukeskuksiin keskittyminen, työelämän ja toimeentulon epävarmuudet sekä ekologinen kestävyyskriisi. Hyvinvointivaltion rahoitus on tiukkaa, ja terveydenhuollossa on osaajapula niin kansallisesti kuin globaalisti. Sitra (Dufva, 2024) tarjoaa ratkaisuksi uuden teknologian ja terveysdatan luotettavaa käyttöä. Myös vuonna 2023 tehtyyn hallitusohjelmaan on kirjattu: ”Selvitetään todennäköisimmin automatisoitavissa olevat sosiaali- ja terveydenhuollon tehtävät sekä niihin liittyvät mahdollisuudet ja riskit.” (Valtioneuvosto, 2023, s. 40).

Yhdysvaltalaisen tutkimuksen (Toscano ja muut, 2020, s. 3166) mukaan lähes puolet lääkäreiden työajasta kului sähköisen potilastietojärjestelmän parissa, muun muassa kirjaamiseen ja lausuntoihin. Konsulttiyhtiö Accenturen (2017, s. 4, 6) analyysin mukaan jopa

20 % sosiaali- ja terveydenhuollon henkilöstön töistä voisi olla siirrettävissä generatiivisen tekoälyn tehtäväksi. Tekoäly voi siis tarjota merkittävää hyötyä tarjoten terveydenhuollon ammattilaisille enemmän aikaa asiakkaiden ja potilaiden parissa. Sanmark ja Sanmark (2024) suosittelevat, että generatiivinen tekoäly tulisi tuoda osaksi suomalaista sosiaali- ja terveydenhuoltoa jo ennen kattavan tieteellisen näytön kertymistä. Heidän mukaansa käyttökohteet tulisi valita teknologian rajoitukset huomioiden ja kattavaa riskienarviointia käyttäen.

Sohail ja muut (2023, s. 17) mainitsevat tekoälyn luotettavuuden kehittämisen olevan yksi merkittävimpiä kehityskohteita. Tekoälyn luotettavuudelle ei ole yhtä yksiselitteistä määritelmää, ja aiheesta on laadittu lukuisia ohjeistuksia ja standardeja (El-Sappagh ja muut, 2023, s. 11 157). Euroopan komission teettämä raportti (2019) on yksi käytetyimmistä lähteistä tekoälyn luotettavuuden arviointiin, vaikka Rathkopf ja Heinrichs (2024, s. 333–339) kritisoivatkin sen hajanaisia, ihmiskeskeisiä oletuksia sekä luotettavuuden käsitteen puutteellista määrittelyä. Raportti määrittelee luotettavan tekoälyn kolmeksi perusedellytykseksi lainmukaisuuden, eettisyyden ja teknisen sekä sosiaalisen luotettavuuden, jotka ovat yhteydessä tekoälyn käyttöympäristöön ja sen vaikutuksiin ihmisiin ja yhteiskuntaan (Euroopan komissio, 2019, s. 6).

Euroopan komission teettämä raportti (2019, s. 10) nostaa luotettavan tekoälyn keskeisiksi eettisiksi periaatteiksi ihmisen itsemääräämisoikeuden kunnioittamisen, vahinkojen välttämisen, oikeudenmukaisuuden ja selitettävyyden. Itsemääräämisoikeus edellyttää, ettei tekoäly saa alistaa tai manipuloida ihmisiä, vaan sen tulee tukea ja vahvistaa ihmisten taitoja. Oikeudenmukaisuus puolestaan kattaa hyötyjen ja haittojen tasapuolisen jakautumisen sekä mahdollisuuden riitauttaa tekoälyn päätökset. Selitettävyys tarkoittaa, että tekoälyjärjestelmien toiminnan ja päätösten tulee olla ymmärrettäviä ja jäljitettäviä, mikä on erityisen tärkeää terveydenhuollossa (Euroopan komissio, 2019, s. 14–15; Sánz ja muut, 2024, s. 7). Käytännössä luotettavuutta arvioidaan Euroopan komission (2019, s. 17) mukaan seitsemällä vaatimuksella, joita ovat ihmisen toimijuus ja valvonta, tekninen luotettavuus ja turvallisuus, yksityisyydensuoja ja datan hallinta, läpinäkyvyys,

monimuotoisuus ja syrjimättömyys, yhteiskunnallinen ja ekologinen hyvinvointi sekä vastuuvelvollisuus.

Terveydenhuollon tekoälyjärjestelmät määritellään suuririskisiksi järjestelmiksi, mikä asettaa tiukat vaatimukset muun muassa riskienhallinnalle ja datan käsittelylle (Euroopan parlamentin ja neuvoston tekoälyasetus, 2024/1689). Tekoälyn käyttöönottoon liittyy merkittäviä haasteita, kuten datan laatuongelmat, epätarkkuudet sekä mahdolliset sosiaaliset ja juridiset seuraukset (El-Sappagh ym., 2023, s. 11 150; Sáez ym., 2024, s. 1–2). Vaikka tekoälyn hyödyntämiseen terveydenhuollossa liittyykin haasteita, ovat tutkijat yhtä mieltä sen potentiaalista parantaa terveydenhuollon laatua ja tehokkuutta, kunhan luotettavuuden varmistamiseksi asetetut vaatimukset täyttyvät ja haasteisiin vastataan systemaattisesti (El-Sappagh ym., 2023, s. 11 276; Sanmark & Sanmark, 2024).

Tutkimuskysymykset työssä ovat: 1. Millaisia haasteita nykytilanteessa on havainnoitu luotettavan tekoälyn suhteen terveydenhuollossa? 2. Millaisia ratkaisuja luotettavuuden parantamiseksi on löydetty? Työn tutkimusmenetelmänä on kuvaileva, narratiivinen kirjallisuuskatsaus. Kirjallisuuskatsauksen ollessa systemaattinen ja täsmällinen menetelmä, jolla voidaan tunnistaa, arvioida ja tiivistää jo julkaistua tutkimusaineistoa tehden siitä johtopäätöksiä (Salminen, 2023, s. 4), soveltuu se erityisen hyvin työn aihealueeseen, josta uutta tutkimustietoa kertyy nopealla tahdilla.

Luvussa kaksi käydään läpi työn teoriapohjaa ja käsitteiden määrittelyä tekoälyn, luotettavan tekoälyn ja terveydenhuollon tekoälyn osalta. Luku kolme sisältää tutkimusmenetelmään liittyvät asiat, ja luvussa neljä paneudutaan tutkimustuloksiin. Luvusta viisi löytyvät pohdinta, johtopäätökset sekä ehdotetut jatkotutkimusaiheet.

2 Tekoäly terveydenhuollossa

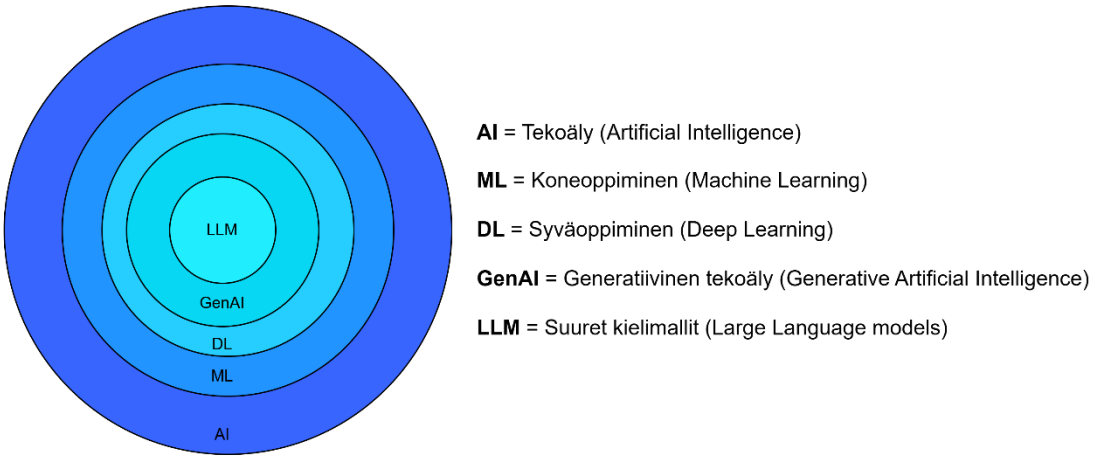
Tässä luvussa määritellään tekoälyn ja luotettavan tekoälyn käsitteet. Tekoälyn määrittäminen ei ole helppoa, sillä siitä ei ole olemassa yhtä yleisesti hyväksyttyä määritelmää (Sheikh ja muut, 2023, s. 15). Aiemman tutkimustiedon pohjalta käydään myös läpi, miten tekoäly terveydenhuollossa näyttäytyy ja kuinka sitä voidaan hyödyntää. Lisäksi tarkastellaan tekoälyn erityispiirteitä ja haasteita terveydenhuollossa.

2.1 Tekoäly käsitteenä

Sheikh ja muiden (2023, s. 15) mukaan laajimman määritelmän mukaan tekoäly rinnastetaan algoritmeihin. Tämä ei kuitenkaan useinkaan ole riittävä lähestymistapa, sillä algoritmeja käytetään myös laajalti tekoälyn ulkopuolella. Tiukimman määritelmän mukaan tekoäly tarkoittaa ihmisille ominaista älykkyyttä, jota tietokone jäljittelee. Tämä tiukka määritelmä kuitenkin puolestaan rajaa monet nykyisistä yksinkertaisemmista sovelluksista tekoälyn käsitteen ulkopuolelle. Tekoälyn yleismääritelmänä toimiikin usein se, että se on teknologiaa, joka mahdollistaa koneelle erilaisten monimutkaisten ihmistaitojen jäljittelyn. (Sheikh ja muut, 2023, s. 15).

Tätä tukee myös Euroopan parlamentin ja neuvoston tekoälyasetuksen (2024/1689, 3. artikla) määritelmä tekoälyjärjestelmästä konepohjaisena järjestelmänä, joka on suunniteltu toimimaan käyttöönoton jälkeen vaihtelevilla autonomian tasoilla ja jossa voi ilmetä mukautuvuutta käyttöönoton jälkeen. Määritelmän mukaan järjestelmä myös päättelee vastaanottamastaan syötteestä eksplisiittisiä tai implisiittisiä tavoitteita varten, miten tuottaa tuotoksia, kuten ennusteita, sisältöä, suosituksia tai päätöksiä, jotka voivat vaikuttaa fyysisiin tai virtuaalisiin ympäristöihin. Asetuksen 12. kohdassa yhdeksi tekoälyjärjestelmien keskeisistä piirteistä mainitaan niiden päättelykyky, jonka tekniikkana toimii esimerkiksi koneoppimismenetelmät. Tekoälyjärjestelmän päättelykyvyn sanotaan olevan laajempi kuin perustietojenkäsittelyn. (Euroopan parlamentin ja neuvoston asetus (EU) 2024/1689, 2024).

Russell ja Norvig (2021, s. 19–22) puolestaan määrittelevät tekoälyn neljän eri lähestymistavan kautta: Turingin testin, ihmismäisen ajattelun, rationaalisen ajattelun ja rationaalisen toiminnan perusteella. Turingin testi mittaa koneen kykyä jäljitellä ihmisen ajattelua kirjallisessa keskustelussa, ja laajennettu eli täydellinen Turingin testi vaatii lisäksi vuorovaikutusta fyysisessä maailmassa. Ihmismäisen ajattelun määritelmän haasteena on, ettei ihmisen kognitiota täysin tunneta, joten kognitiotiede yhdistää tekoälyn mallit ja psykologian menetelmät luodakseen testattavia teorioita ihmismielestä. Rationaalinen ajattelu puolestaan perustuu logiikkaan ja todennäköisyyksiin, kun taas rationaalinen käyttäytyminen yleistää älykästä toimintaa. Tässä yhteydessä puhutaan rationaalisista agenteista, jotka toimivat saavuttaakseen parhaan mahdollisen lopputuloksen tai epävarmuuden vallitessa parhaan odotettavissa olevan lopputuloksen. Rationaalisen agentin lähestymistapaa on hyödynnetty paljon tekoälyn olemassaolon aikana. (Russell & Norvig, 2021, s. 21–22)



Kuvio 1. Tekoälyn hierarkia ja tasot (mukailtuna Sanmark & Sanmark, 2024)

Tekoäly on kattotermi, jonka alle mahtuu paljon. Sanmarkin ja Sanmarkin (2024) kuvion mukaelmasta (kuvio 1) näkyy termistön hierarkia. Koneoppimisella tarkoitetaan tekoälyn haaraa, jossa tietokonejärjestelmät oppivat ja parantavat suorituskykyään kokemuksen perusteella ilman ohjelmointia. Syväoppiminen puolestaan on koneoppimisen menetelmä, joka käyttää monikerroksisia hermoverkkoja oppiakseen monimutkaista dataa tehokkaasti ja laajasti. Generatiivinen tekoäly taas viittaa tekoälysovelluksiin, jotka

kykenevät luomaan uutta sisältöä, kuten tekstiä tai kuvia, oppimansa datan pohjalta. Suuret kielimallit ovat tekoälyjärjestelmiä, jotka on koulutettu ymmärtämään ja tuottamaan kieltä, mikä mahdollistaa monipuoliset kielelliset tehtävät. (Sanmark & Sanmark, 2024)

2.2 Tekoälyn luotettavuus

El-Sappaghin ja muiden (2023, s. 11 157) mukaan tekoälyn luotettavuudelle ei ole olemassa yhtä yksittäistä määritelmää. Jo heidän artikkelinsa julkaisun aikaan on ollut olemassa yli 60 korkeatasoista ohjeistusta ja standardia tekoälyn luotettavuuden ja etiikan arviointiin ja tarkkailuun. He viittaavat Jacoviin ja muihin, joiden mukaan Euroopan komission teettämä raportti (ALTAI) on kaikkein käyttökelpoisin ohjeistus tekoälyjärjestelmien luotettavuuden mittaamiseen.

Rathkopf ja Heinrichs (2024, s. 333-334, 338-339) puolestaan kritisoivat koko luotettavan tekoälyn käsitettä. Heidän mukaansa tavoite luotettavasta tekoälystä voi olla vaarallinen, koska se pakottaa meidät omaksumaan kyseenalaisia asenteita, ja perustuu hajanaisiin ihmiskeskeisiin oletuksiin. He eivät kuitenkaan kannata myöskään epäluottamusta tekoälymalleja kohtaan, vaan korostavat huolellista ja tarkkaa virheanalyysiä. He myös kritisoivat Euroopan komission teettämää raporttia alikehittyneeksi, sillä se ei tarjoa heidän mukaansa luotettavuuden määritelmää, tai perustele miksi luotettavuus on sopiva perusta tekoälyn etiikalle. Heidän mukaansa raportti keskittyy teknisiin ja lainsäädännöllisiin strategioihin, mutta ei huomioi riittävästi koneoppimisen luokittelupäätösten eroja ihmisen tekemiin arvioihin verrattuna. Saamastaan kritiikistä huolimatta, Euroopan komission teettämää raporttia käytetään tässä työssä luotettavan tekoälyn kehyksenä, sillä se on asiayhteydessään laajasti käytetty ja viitattu dokumentti.

Euroopan komission teettämän raportin (2019, s. 6) mukaan luotettavalla tekoälyllä on kolme edellytystä, joiden tulisi täyttyä järjestelmän elinkaaren kaikissa vaiheissa: lainmukaisuus, eettisyys ja luotettavuus. Lainmukaisuus tarkoittaa kaikkien sovellettavien lakien

ja asetusten noudattamista, eettisyys eettisten periaatteiden ja arvojen noudattamisen varmistamista, luotettavuus sekä teknistä että sosiaalista luotettavuutta, ja tahattoman haitan ja vahingon välttämistä. Tekoälyn eettisyys ja luotettavuus liittyvätkin läheisesti toisiinsa, ja täydentävät toisiaan. Luotettavuuden sanotaan olevan ennakkoedellytys sille, että ihmiset ja yhteiskunnat kehittävät, ottavat käyttöön ja käyttävät tekoälyjärjestelmiä. Tämä luottamus ei koske vain teknologian sisäisiä ominaisuuksia, vaan myös sitä, millaisessa ympäristössä tekoälyä käytetään ja miten se vaikuttaa ihmisiin, organisaatioihin ja yhteiskuntaan. Luotettavan lähestymistavan sanotaan myös olevan vastuullisen kilpailukyvyn edellytys. (Euroopan komissio 2019, s. 6). Sohail ja muut (2023, s. 19) mainitsevat ihmiskeskeisen suunnittelun tukevan näitä edellytyksiä.

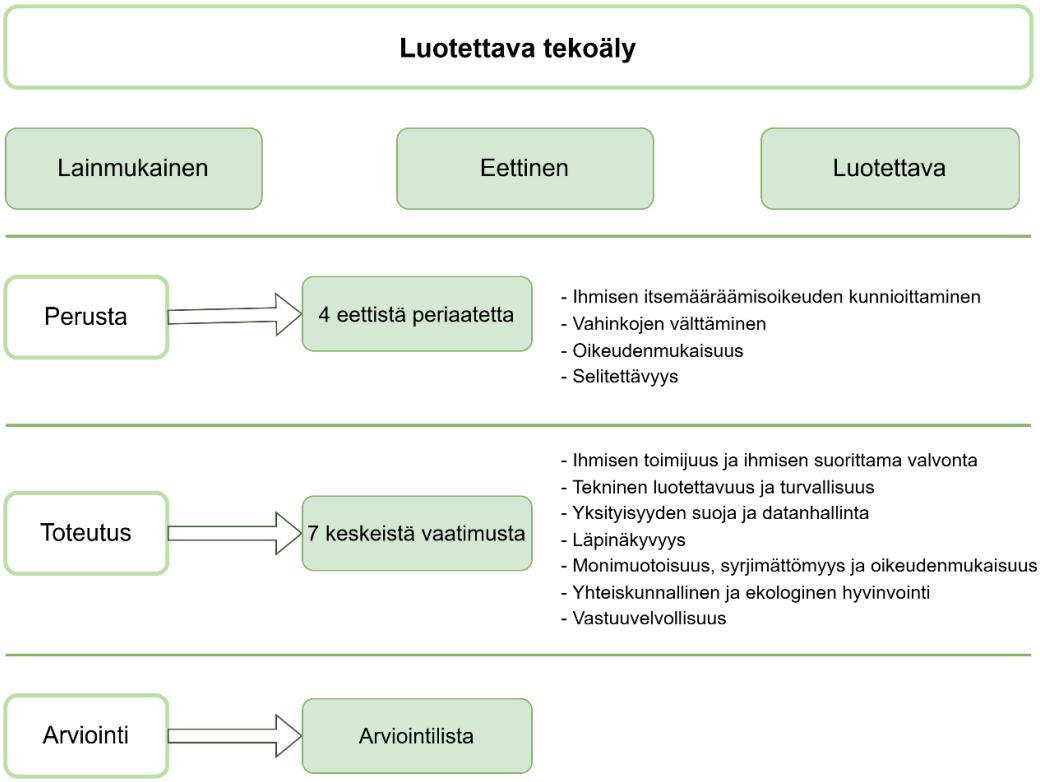
2.2.1 Luotettavan tekoälyn periaatteet

Euroopan komission (2019, s. 10) ALTAI:n kokonaiskehystä voi mukailtuna tarkastella kuviosta 2. Periaatteita, toteutusta ja arviointia avataan laajemmin tässä ja seuraavassa alaluvussa. Luotettavan tekoälyn perustalla varmistetaan neljän eettisen periaatteen noudattaminen: ihmisen itsemääräämisoikeuden kunnioittaminen, vahinkojen välttäminen, oikeudenmukaisuus ja selitettävyys.

Euroopan komission raportin (2019, s. 10) mukaan tekoälyjärjestelmiä käyttävien ihmisten on voitava säilyttää täysimääräinen ja tehokas itsemääräämisoikeus, osallistuen demokraattiseen prosessiin. Tekoälyjärjestelmät siis eivät saisi perusteettomasti alistaa, pakottaa, johtaa harhaan, manipuloida, ehdollistaa tai holhota ihmisiä. Sen sijaan niiden tulisi lisätä, täydentää ja vahvistaa ihmisen kognitiivisia, sosiaalisia ja kulttuurisia taitoja. Tekoälyjärjestelmät eivät myöskään saa aiheuttaa eivätkä pahentaa vahinkoja, tai muutoin vaikuttaa haitallisesti ihmisiin. Tämä tarkoittaa sekä ihmisarvon, että henkisen ja ruumiillisen koskemattomuuden suojelua. Järjestelmien on siis oltava turvallisia.

Oikeudenmukaisuus kattaa sekä aineellisen että menettelyllisen ulottuvuuden. Aineellinen oikeudenmukaisuus tarkoittaa sitoutumista, ja hyötyjen sekä kustannusten

tasapuolista ja oikeudenmukaista jakautumista, sekä yksilöihin ja ryhmiin kohdistuvan epäoikeudenmukaisen puolueellisuuden, syrjinnän ja leimaamisen ennaltaehkäisyä. Menettelyllinen oikeudenmukaisuus edellyttää mahdollisuutta riitauttaa tekoälyjärjestelmien ja niitä käyttävien ihmisten tekemät päätökset, sekä tehokasta muutoksenhakua päätöksiin. Jos nämä vääristymät vältetään, tekoälyjärjestelmät voivat jopa lisätä yhteiskunnallista oikeudenmukaisuutta. (Euroopan komissio, 2019, s. 14–15). Selitettävyyden periaate puolestaan tarkoittaa Euroopan komission raportin (2019, s. 14–15) mukaan sitä, että prosessien on oltava avoimia, tekoälyjärjestelmien kapasiteetti ja tarkoitus on ilmoitettava avoimesti, ja mahdollisuuksien mukaan päätökset on pystyttävä selittämään. Ilman näitä tietoja päätöstä ei voi riitauttaa asianmukaisesti.



Kuvio 2. Luotettavan tekoälyn kehys (mukailtuna Euroopan komissio 2019, s. 10)

Tilanteita, joissa ei ole mahdollista selittää miten jokin malli on tuottanut tietyn tuloksen tai päätöksen, kutsutaan ”mustan laatikon” (black box) algoritmeiksi. Nämä tilanteet vaativat erityishuomiota, ja edellyttävät usein muita selitettävyyttä koskevia toimenpiteitä,

kuten jäljitettävyys, tarkastettavuus ja avoin tiedotus järjestelmän kyvyistä. Selitettävyyden tasovaatimus riippuu siitä, missä kontekstissa tekoälyjärjestelmää käytetään, ja kuinka vakavat seuraukset virheellisistä tai epätarkoista tuloksista seuraa. Selitettävyydellä on suuri merkitys terveydenhuollon tekoälyjärjestelmissä (Euroopan komissio, 2019, s. 15), sillä mustan laatikon ilmiö on tyypillinen esimerkiksi lääketieteellisessä kuvantamisessa (Sánz ja muut, 2024, s. 7).

2.2.2 Luotettavan tekoälyn toteutus ja arviointi

Eettisten periaatteiden toteutumista arvioidaan seitsemällä keskeisellä vaatimuksella, jotka ovat 1. ihmisen toimijuus ja ihmisen suorittama valvonta, 2. tekninen luotettavuus ja turvallisuus, 3. yksityisyyden suoja ja datan hallinta, 4. läpinäkyvyys, 5. monimuotoisuus, syrjimättömyys ja oikeudenmukaisuus, 6. yhteiskunnallinen ja ekologinen hyvinvointi, 7. vastuuvelvollisuus. Näitä vaatimuksia käytetään tekoälyjärjestelmän elinkaaren kaikissa vaiheissa ja kaikkien sidosryhmien osalta. (Euroopan komissio, 2019, s. 17)

Euroopan komission raportin (2019, s. 19) mukaan ihmisen toimijuus edellyttää käyttäjille on mahdollisuutta tehdä itsenäisiä, tietoon perustuvia päätöksiä tekoälyjärjestelmiin liittyen. Ihmisille tulee tarjota tiedot ja välineet, joiden avulla he voivat ymmärtää, olla vuorovaikutuksessa sekä kohtuullisella tasolla arvioida ja kyseenalaistaa järjestelmien päätöksiä. Ihmisen suorittama valvonta varmistaa, ettei järjestelmä heikennä ihmisten itsenäistä päätäntävaltaa tai aiheuta muuta haittaa. Joustava ihmisen ja tietokoneen välinen vuorovaikutus sekä ihmisen osallistumista edellyttävät mekanismit tukevat selitettävyyttä ja luotettavuutta, sekä takaavat perusoikeudet. (Sánz ja muut, 2024, s. 7; Sohail ja muut, 2023, s. 19).

Tekninen luotettavuus ja turvallisuus tarkoittaa Euroopan komission raportin (2019, s. 20) mukaan vastustuskykyä hyökkäyksiä vastaan, varasuunnitelmaa, tarkkuutta ja toistettavuutta, jolloin järjestelmä käyttäytyy ennakoidulla tavalla sekä minimoi tahattomat ja odottamattomat vahingot, estäen kohtuuttomat vahingot. Tämä on merkittävä tekijä

terveydenhuollon tekoälyn onnistumisessa (Sánz ja muut, 2024, s. 10). Tarkkuus puolestaan liittyy tekoälyjärjestelmän kykyyn tehdä oikeita arvioita, kuten luokitella tietoja oikein, tehdä paikkansapitäviä ennusteita, suosituksia ja päätöksiä. Jos epätarkkuuksia ei voida täysin välttää, tulee järjestelmän ilmoittaa virheiden todennäköisyys. Myös tulosten toistettavuus on keskeistä, järjestelmän on toimittava asianmukaisesti eri syötetiedoilla ja eri tilanteissa, sekä käyttäydyttävä samalla tavoin saman operaation toistuessa samoissa olosuhteissa. (Euroopan komissio, 2019, s. 20)

Yksityisyydensuoja kattaa käyttäjän antamat tiedot ja tekoälyjärjestelmän kanssa vuorovaikutuksessa tuotetut tiedot. Se vaatii asianmukaista datan hallintaa eli tiedon laadun, eheyden ja saatavuuden varmistamista. Datan laadulla on merkittävä vaikutus tekoälyjärjestelmien suorituskykyyn. (Euroopan komissio, 2019, s. 21) Koulutusdatan tulee olla monipuolista, edustaa populaatiota sekä huolellisesti kuratoitua ja suodatettua, jotta vääristynyt ja ongelmallinen data voidaan poistaa jo ennen käyttöä (Sohail ja muut, 2023, s. 18). Tiedon eheydestä on huolehdittava, jotta järjestelmään ei päästä syöttämään haitallisia, järjestelmän käyttäytymistä muuttavia tietoja. Testaaminen ja dokumentointi on välttämätöntä jokaisessa vaiheessa. (Euroopan komissio, 2019, s. 21) El-Sappaghin ja muiden (2023, s. 11 276) mukaan data jaetaan usein harjoitus- ja validointidataan vasta sen jälkeen, kun sille oli suoritettu esivalmisteluvaiheet. Tämä johti informaatiovuotoon, joka voi aiheuttaa sen, että malli tarjoaa optimistisia ja epätosia vastauksia. Tiedon saatavuutta on säänneltävä erityisesti henkilötietoja käsittelevissä organisaatioissa (Euroopan komissio, 2019, s. 21), eli ketkä ja missä olosuhteissa tietoon pääsevät käsiksi.

Euroopan komission raportin (2019, s. 21–22) mukaan läpinäkyvyys kattaa tietojen, järjestelmän ja liiketoimintamallien läpinäkyvyyden, tarkoittaen jäljitettävyyttä, selitettävyyttä ja tiedotusta. Päätökseen vaikuttavat tekijät kuten tietojen kerääminen, käytetyt tunnisteet sekä algoritmit tulee dokumentoida mahdollisimman hyvin. Jäljitettävyys auttaa tunnistaa syyt tekoälyjärjestelmän virheelliseen päätökseen, ja siten ennaltaehkäisee virheitä jatkossa. Selitettävyys tarkoittaa kykyä selittää niin tekoälyjärjestelmän tekniset prosessit kuin siihen liittyvät ihmisen päätökset. Tekninen selitettävyys vaatii, että

ihmiset voivat ymmärtää ja jäljittää järjestelmän tekemät päätökset. (Euroopan komissio, 2019, s. 22). Raportti (2019, s. 22) sekä El-Sappagh ja muut (2023, s. 11 276) nostavat esiin luotettavan tekoälyn eri osa-alueiden ristiriidat. Esimerkiksi selitettävyyden parantuessa tarkkuus voi heikentyä, ja tarkkuuden lisääntyessä taas selitettävyys voi heikentyä. Vastaavanlaiset dilemmat koskevat useita eri luotettavan tekoälyn osa-alueita. Sanmarkin ja Sanmarkin (2024) mukaan tietosuoja- ja luottamusriskit liittyvät merkittävästi siihen, ettei kielimallien lähdekoodiin tai koulutusaineistoihin ole näkyvyyttä, kun markkinoita hallitsevat kaupallisten toimijoiden kielimallit. Gargin ja muiden (2025) mukaan syväoppiminen on mullistanut lääketieteellisen kuvantamisen tarjoten siihen tehokkaita työkaluja, mutta mallit kuitenkin kärsivät läpinäkymättömyydestä.

Tekoälyjärjestelmät eivät saa esiintyä ihmisinä, vaan käyttäjillä on oikeus saada tietää olevansa tekemisissä tekoälyn kanssa (Euroopan komissio, 2019, s. 22). Rathkopf ja Heinrichs (2024, s. 342) painottavat, että potilaiden on tiedettävä mihin he sitoutuvat, kun tekoälyä käytetään lääketieteessä. Vain siten he voivat itse päättää, haluavatko hyödyntää kyseistä teknologiaa vai eivät. Jos asia salataan, näkevät Rathkopf ja Heinrichs, että kehittäjät, ylläpitäjät ja käyttäjät ovat vastuullisia, mikäli potilaille aiheutuu haittaa, jota he eivät salaamisen vuoksi olleet voineet ennakoida.

Luotettavan tekoälyn on mahdollistettava osallistavuus ja monimuotoisuus tekoälyjärjestelmän elinkaaren kaikissa vaiheissa. Yhtäläiset mahdollisuudet tulee varmistaa esimerkiksi osallistavilla suunnitteluprosesseilla ja tasapuolisella kohtelulla. Tekoälyjärjestelmän kouluttamiseen ja toimintaan käytetyt tietokokonaisuudet voivat sisältää tahattomia vääristymiä ja epätäydellisyyksiä, joiden jatkuminen järjestelmän sisällä voi johtaa tiettyjen ryhmien tai henkilöiden syrjintään tai pahentaa ennakkoluuloja ja syrjäytymistä. Yhteiskunnallisesti rakentuneisiin vääristymiin, epätarkkuuksiin ja puutteellisuuksiin on puututtava jo ennen kuin tekoälyjärjestelmä koulutetaan tietyllä tietokokonaisuudella. Haittaa voidaan aiheuttaa myös tarkoituksellisesti esimerkiksi kuluttajasuuntauksien hyödyntämisen tai vilpillisen kilpailun avulla. Käytettävien tekoälyjärjestelmien tulee

myös olla käyttäjälähtöisiä ja suunnittelussa tulee noudattaa esteettömyysstandardeja. (Euroopan komissio, 2019, s. 22–23)

Tekoälyjärjestelmillä on vastuu yhteiskunnallisesta ja ekologisesta hyvinvoinnista, ja niiden ympäristöystävällisyys on varmistettava prosessin kaikissa vaiheissa, esimerkiksi resurssien ja energiankulutuksen käyttö huomioiden. Euroopan komission raportti (2019, s. 23) nostaa esiin järjestelmien sosiaaliset vaikutukset, joita jatkuva altistuminen vuorovaikutteisille tekoälyjärjestelmille voi aiheuttaa. Se voi muuttaa käsitystämme sosiaalisista suhteista ja toimijuudesta. Vaikutuksia onkin seurattava ja arvioitava niin yksilön kuin yhteiskunnan näkökulmasta.

Vastuuvelvollisuus korostuu tekoälyjärjestelmän koko elinkaaren ajan, kattaen tarkastettavuuden eli algoritmien, tiedon ja prosessien arvioinnin. Haitalliset vaikutukset tulee tunnistaa, arvioida ja minimoida, vaikutusten arviointi ennen kehittämistä ja käyttöönottoa sekä niiden aikana auttaa vähentämään riskejä. Kaikkien läpikäytyjen vaatimusten välillä voi syntyä jännitteitä, jotka johtavat kompromisseihin. Niiden on oltava rationaalisia ja järjestelmällisiä, ja jos hyväksyttäviä kompromisseja ei löydy, tulee järjestelmän kehittäminen ja käyttö siinä muodossaan lopettaa. Järjestelmiin tulee myös sisällyttää helppokäyttöisiä muutoksenhakumekanismeja virhetilanteiden varalta. (Euroopan komissio, 2019, s. 24.)

Euroopan komission raportti (2019, s. 29–40) tarjoaa myös arviointilistan, jota käytetään tekoälyn eettisten periaatteiden ja vaatimusten konkretisoimisessa. Arviointilista on aina räätälöitävä konkreettisiin käyttötilanteihin ja konteksteihin. Tekoälyjärjestelmää suunnitellessa, käyttöönottaessa tai käytettäessä on syytä laatia kysymyslista arvioinnin tueksi. Tällainen arviointilista ei kuitenkaan koskaan ole tyhjentävä, vaan luotettavuuden varmistaminen on jatkuva prosessi. (Euroopan komissio, 2019, s. 29.)

2.3 Tekoäly ja sen hyödyntäminen ja mahdollisuudet terveydenhuollossa

Euroopan parlamentin ja neuvoston tekoälyasetuksen (2024/1689, liite 3) perusteella terveydenhuollossa käytettävät tekoälyjärjestelmät ovat suuririskisiä, sillä suuririskisiksi tekoälyjärjestelmiksi määritellään muun muassa sellaiset, joita käytetään kriittisen infrastruktuurin ylläpitoon. Eli tekoälyjärjestelmät, jotka on tarkoitettu käytettäväksi turvakomponentteina esimerkiksi kriittisen digitaalisen infrastruktuurin hallinnassa ja toiminnassa. Turvakomponentit ovat EU:n tekoälyasetuksen 3. artiklan 14. kohdan mukaan tuotteen tai tekoälyjärjestelmän komponentteja, joka toteuttavat laitteen tai tekoälyjärjestelmän turvallisuustoimintoja, ja joiden vikaantuminen voi vaarantaa ihmisten terveyden ja turvallisuuden. Edelleen liitteen 3. mukaan suuririskisiksi tekoälyjärjestelmiksi määritellään biometrisiä tunnisteita, kuten biometrisiä etätunnistusjärjestelmiä ja tekoälyjärjestelmiä käyttävät, jotka on tarkoitettu käytettäväksi biometriseen luokitteluun arkaluonteisten tai suojattujen ominaisuuksien tai ominaispiirteiden mukaisesti.

Lääkinnälliset laitteet määritellään EU:ssa tuotteiksi, jotka kuuluvat unionin yhdenmukaistamislainsäädännön soveltamisalaan. Sillä perusteella kaikki lääkinnälliset laitteet määritellään suuririskisiin tekoälyjärjestelmiin kuuluviksi, jos niissä käytetään tekoälyominaisuuksia. Tekoälyasetuksen artikloissa 9–14 määritellään ja säädellään suuririskisten tekoälyjärjestelmiin liittyen pakollisesta riskienhallintajärjestelmästä, datasta ja datanhallinnasta, teknisestä dokumentaatiosta, tietojen säilyttämisestä, avoimuudesta ja tietojen antamisesta käyttöönottajille, ihmisen suorittamasta valvonnasta, sekä tarkkuudesta, vakaudesta ja kyberturvallisuudesta.

Euroopan komission teettämän raportin (2019, s. 41–42) mukaan tekoälyteknologiaa voidaan käyttää muuttamaan hoitoa älykkäämmäksi ja kohdennetummaksi, sekä ehkäisemään hengenvaarallisia sairauksia. Terveydenhuollon ammattilaisilla on mahdollisuus tehdä tarkempia ja yksityiskohtaisempia analyyseja potilaan terveystiedoista jo ennen hänen sairastumistaan, sekä antaa räätälöityä ehkäisevää hoitoa. Tekoäly ja robotiikka voivat olla hoitotyötekijöitä avustavia ja tukevia välineitä, joilla voidaan muun muassa seurata potilaiden tilaa reaaliajassa. Raportti jatkuu maininnalla luotettavan tekoälyn

tarjoamista mahdollisuuksista myös tutkia ja havaita yleisiä kehityssuuntauksia terveydenhuolto- ja hoitosektorilla, jolloin voidaan edistää sairauksien varhaisempaa havaitsemista, tehostaa lääkkeiden kehittämistä ja kohdentaa hoitoja tehokkaammin. Bærøe ja muut (2020, s. 257) mainitsevat yllä lueteltujen lisäksi myös komplikaatioiden hallinnan ja lieventämisen kohteena, jossa tekoälyä jo käytetään terveydenhuollossa.

Generatiivisen tekoälyn hyödyntämisestä terveydenhuollossa on niin kansallisesti kuin kansainvälisesti vielä rajallisesti kokemuksia ja tieteellistä tietoa. Sanmarkin ja Sanmarkin (2024) mukaan on kuitenkin esitetty, että generatiivinen tekoäly pystyisi esimerkiksi tekemään yhteenvetoja ammattilaiselle keskeisistä potilastiedoista ja sairauskertomusmerkinnät nauhoitetun vastaanoton perusteella, laatimaan lausuntoja ja räätälöityjä potilasohjeita, hakemaan tietoa tietokannoista, auttamaan riskipotilaiden tunnistamisessa, tehostamaan potilasviestintää sekä parantamaan tiedolla johtamista. Sanmark ja Sanmark (2024) viittaavat useisiin tutkimuksiin, joiden mukaan generatiivisella tekoälyllä on merkittävää potentiaalia sairaalahoidon keston, kuolleisuuden tai sairaalaan paluun ennustamisessa, dokumenttien luomisessa sekä radiologisten kuvantamistutkimusten tulkinnassa. Näistä generatiivinen tekoäly oli tutkimuksissa suoriutunut vähintään yhtä hyvin tai jopa paremmin kuin nykyisin käytettävät menetelmät. Tutkijat huomauttivat, että kielimallit suoriutuisivat lääketieteellisistä tehtävistä paremmin, jos ne olisi koulutettu lääketieteellisellä datalla. Kahdessa tutkimuksessa oli käytetty lääketieteelliseen toimintaan tarkoitettua kielimallia, jotka suoriutuivat ihmisen veroisesti tai aiempia menetelmiä paremmin. Myös promptin eli kielimallille annettavan komennon optimointi parantaa kielimallin tarkkuutta tehtävissä. (Sanmark & Sanmark, 2024)

Hoidon tarpeen arviointiohjelmistot (Clinical Decision Support Systems, CDSS) ovat puolestaan osoittautuneet hyödylliseksi esimerkiksi erityyppisten päänsärkyjen diagnosoinnissa ja rytmihäiriöiden tunnistamisessa. Henkilökohtaiset digitukijärjestelmät myös parantavat merkittävästi maallikoiden antaman ensiavun laatua. (Liu ja muut, 2020, s. 43-44). El-Sappagh ja muut (2023, s. 11150) viittaavat Sappaghiin ja muihin sekä Abuhmediin ja muihin, joiden mukaan hoidon tarpeen arviointiohjelmistot voivat vähentää

päätöksenteon virheitä, tehostaa ja parantaa hoidon laatua sekä optimoida henkilökohtaisen lääkityksen aikataulutusta. Euroopan komission lisäksi Yhdysvaltain lääketieteellisen informatiikan yhdistys (American Medical Informatics Association) on antanut suosituksensa tekoälyyn pohjautuvan hoidon tarpeen arviointiohjelmiston käyttöön. Yhdistys nostaa esiin huolensa mahdollisista sääntelyongelmista, jotka tulee huomioida: suunnittelu- ja kehitysvaiheen läpinäkyvyys, toteutus viestinnän standardeja ja uudelleenkoulutuskriteerejä noudattaen, käyttöpaikassa toteutettava arviointi ja testaus sekä jatkuva seuranta sisältäen järjestelmän ylläpidon sekä käyttäjäkoulutuksen. (Sánz ja muut, 2019, s. 7). El-Sappagh ja muut (2023, s. 11150) puolestaan viittaavat He ja muihin, joiden mukaan hoidon tarpeen arviointiohjelmiston potilaskeskeiset päätökset tulee pohjata täydelliseen potilasdataan, joka siis sisältää demografiset tiedot, kuvantamistulokset, laboratoriotulokset, genetiikan ja muut sähköisestä potilasrekisteristä kerätyt tiedot.

Terveydenhuollon tekoälymallit ovat rakenteeltaan monitahoisia. Rathkopf ja Heinrichs (2024, s. 342–343) viittaavat Finlaysoniin ja muihin, joiden mukaan useimmat terveydenhuollon neuroverkkomallit on kehitetty alun perin tekoäly-yhtiöissä tai tutkimusryhmissä yleisiin luokittelutehtäviin, josta ne on sitten myöhemmin mukautettu terveydenhuollon sovelluksiksi erillisten itsenäisten ryhmien toimesta. Tämä hajautettu kehitysprosessi altistaa ne monen osapuolen vastuuongelmaan (many hands problem), jossa virhevastuuta ei voida kohdistaa yksittäiseen tahoon.

Liu ja muut kertovat (2020, s. 43) satunnaismetsän (random forest) olevan yksi koneoppimisen tehokkaimmista algoritmeista, ja sitä on hyödynnetty myös lääketieteessä erityisesti sairauksien ennakoinnissa. Neuroverkkojen avulla voi esimerkiksi ennustaa sädehoidon tehokkuutta. Älykkäitä robotteja taas on alettu käyttää leikkauksissa jo 1980-luvulla. Yksi uusimmista ja laajenevasti robottikirurgiassa käytössä olevista innovaatioista on jatkuvaliikkeiset robotit, joiden ennakoidaan tulevaisuudessa tekevän yhä isomman osan leikkauksista. Kuvantunnistusteknologiaa taas käytetään niin kuvantamisessa kuin sairauksien varhaisessa tunnistuksessa.

Sáezin ja muiden (2024, s. 2) mukaan terveydenhuollon koneoppimisessa uusi tieto opitaan laskennallisesti koulutusvaiheen datasta, joka on yleensä peräisin kohorttitutkimuksista tai sähköisistä potilasrekistereistä. Tätä tietoa käytetään avustamaan uusien tapausten päätöksenteossa diagnoosi- ja ennustusvaiheessa. Tässä vaiheessa hoidon tarpeen arviointiohjelmisto voi hakea syötteensä sähköisestä potilasrekisteristä tai manuaalisesta syötteestä. Mahdolliset dataan liittyvät ongelmat terveydenhuollon tekoälyssä voivatkin ilmetä sekä koulutus- että päätöksentekovaiheessa. (Sáez ja muut, 2024, s. 2).

2.4 Tekoälyn käytön erityispiirteet ja haasteet terveydenhuollossa

Tekoälyn käyttö terveydenhuollossa ei ole suoraviivaista, sillä siihen liittyy riskejä, jotka voivat johtaa haitallisiin virheisiin, ja aiheuttaa sosiaalisia, laillisia ja taloudellisia seurauksia (El-Sappagh ja muut, 2023, s. 11150). Esimerkiksi hoidon tarpeen arviointiohjelmiston reaaliaikaisessa käytössä on esiintynyt haasteita. Tähän ovat vaikuttaneet El-Sappaghin ja muiden (2023, s. 11150) viittaaman Jacobsin ja muiden listaamat kehitysprosessin tiukkuuden ja kattavuuden puute, järjestelmän tehokkuuden ja käytettävyyden riittämätön arviointi sekä ennustavien tekoälyjärjestelmien luotettavuuden ja vastuullisuuden puutteellinen arviointi. Tämän hetken metodit terveydenhuollon tekoälyn kehittämisessä ovatkin siiloutuneet. He viittaavat myös Ji ja muihin, joiden mukaan mallit on rakennettu suppealla skaalalla laboratorioissa, reaalimaailman asetuksista eristettynä, sekä ilman ison sähköisen potilastietorekisterin ekosysteemiä. Tämä saattaa johtaa siihen, että järjestelmät tuottavat enemmän haittaa kuin hyötyä. (El-Sappagh ja muut, 2023, s. 11151)

Sáezin ja muiden (2024, s. 1) mukaan epätarkkuus, vaihtelevuus ja vääristymät käytännön dataympäristöissämme aiheuttavat merkittäviä haasteita terveydenhuollon tekoälyn kehittämisessä, rutiininomaisessa kliinisessä käytössä sekä lainsäädännöllisessä kehyksessä. Nämä haasteet tarkoittavat esimerkiksi vaihtelevaa informaatiota eri asiayhteyksissä ja eri aikoina, aliedustettuihin ryhmiin liittyviä vääristymiä sekä epätarkkuutta, joka johtuu joko puuttuvasta tai päällekkäisestä informaatiosta, tai tiedon

laatuongelmista. Tästä esimerkkinä toimii El-Sappaghin ja muiden (2023, s. 11 150) mainitsema Thiebesin ja muiden tutkimus, jonka mukaan epävarmuutta esiintyy hoidon tarpeen arviointijärjestelmän jokaisessa vaiheessa: 1) potilas epäonnistuu oireidensa ja tilansa kuvailussa, 2) lääkäri ei tulkitse oikein havaintojaan, 3) laboratoriotulokset sisältävät jonkin tason virheitä tai epätarkkuuksia ja 4) lääketieteellinen data ei ole yhdenmukaista. Sáez ja muut (2024, s. 2) viittaavat Gianfrancescoon, joka on maininnut vinoutuneeseen dataan perustuvan koneoppimismallin mahdollisista negatiivisista vaikutuksista sosioekonomisiin terveyseroihin väestössä.

Terveydenhuollon data on siis epätäydellistä, puutteellista ja vaihtelevuudelle altista. Sáez ja muut (2024, s. 2) viittaavat Cheniin ja Aschiin, joiden mukaan pelkkään menneeseen dataan pohjaava tekoälyjärjestelmä voi pienentää hyötyä ja yleistettävyyttä. Pienet nykyiset muuttujat voivat aiheuttaa perhosefektin tulevaisuudessa. Siksi tekoälyn tulisi olla vastustuskykyinen tällaiselle datalle sen sijaan, että dataa muokattaisiin keinotekoisesti. Keinoja vastustuskyvyn parantamiseen ovat datan laadunhallinta, jatkuva oppiminen, mallien siirrettävyys, perusmallit, keskusteleva tekoäly, ihmisen ja tietokoneen vuorovaikutus sekä sääntely. Näiden keinojen avulla esimerkiksi uuden sukupolven hoidon tarpeen arviointiohjelmistot mahdollistuvat, ja tekoälypohjaisen terveydenhuollon luotettavuus ja luottamus paranevat. (Sánz ja muut, 2024, s. 10; El-Sappagh ja muut, 2023, s. 11 150)

Vaikka tekoälyllä on merkittävä rooli terveydenhuollon kehittämisessä, on sen yllä El-Sappaghin ja muiden (2023, s. 11 166) mukaan edelleen huolia myös tekniikan kestävyydestä terveydenhuollon perspektiivissä. Huolet liittyvät lähinnä turvallisuuteen ja yksityisyydensuojaan tietoturvahyökkäysten yhteydessä. Teknisen kestävyyden saavuttamiseen vaaditaan kolmen pääongelman ratkaisua: turvallisuus ja yksityisyys, tarkkuus sekä luotettavuus ja toistettavuus. Rathkopfin ja Heinrichsin (2024, s. 340) mukaan terveydenhuollon tekoäly on lisääntyvissä määrin erityisen altis haitallisille hyökkäyksille, koska lääketieteellinen data on pirstaloitunutta ja sairaalainfrastruktuurin päivittäminen hidasta ja vaikeaa. Osa näistä hyökkäyksistä voi aiheuttaa outoja ja odottamattomia

virheitä. Rathkopf ja Heinrichs (2024, s. 339) määrittelevät tekoälyn oudot virheet virheiksi, jotka johtuvat ihmisille huomaamattomista syötetietojen häiriöistä ja jotka vaikuttavat ihmisiin merkittävästi, jos virheestä olisi tiedossa.

Rathkopfin ja Heinrichsin (2024, s. 340) mukaan outojen virheiden riski terveydenhuollon tekoälyssä ei ole merkittävä kaksijakoisessa päätöksenteossa, mutta monimutkaisemmissa luokitteluongelmissa, joissa datapisteissä on useita muuttujia, radikaalien virheiden riski kasvaa. Esimerkiksi luonnollisen kielen prosessointia sähköisissä potilasjärjestelmissä käytettäessä radikaaleja virheitä voidaan odottaa. Jos käytettävissä olisi teoria niiden syötteiden tietomalleista, jotka tyypillisesti aiheuttavat outoja virheitä, voitaisiin tietynlaisia virheitä pyrkiä ennakoimaan ja siten vähentämään myös niihin liittyviä riskejä. Siitä huolimatta, että tutkimustieto puolustusstrategioista haitallisia hyökkäyksiä vastaan on kasvanut, ei yleispätevää algoritmista ratkaisua ongelmaan kuitenkaan vielä ole. (Rathkopf & Heinrichs, 2024, s. 340)

El-Sappagh ja muut (2023, s. 11 276) suosittelevat ottamaan tekoälyn elinkaaren kaikissa vaiheissa sekä potilaat että asiantuntijat mukaan arvioimaan tekoälymallin kestävyyttä, oikeudenmukaisuutta ja selitettävyyttä lääketieteen näkökulmasta. He myös suosittelevat perinteisen koneoppimisen elinkaaren laajentamista sisältämään myös luotettavan tekoälyn vaatimukset. Tällä hetkellä elinkaari keskittyy vain mallin tehokkuuteen ja suorituskykyyn, joka ei ole riittävää kriittisiin elintoimintoihin liittyvissä sovelluksissa. Ongelmana on myös se, että tekoälytutkimuksella ei ole tällä hetkellä merkittävää vaikutusta terveydenhuollon toimintaympäristössä, sillä potilaat ja lääketieteen ammattilaiset eivät luota tekoälypohjaisiin päätöksiin. Muun muassa mallien vakaus, varmuus, yleistettävyys, turvallisuus ja yksityisyys tulee ottaa huomioon, jotta tutkimustietoa voidaan hyödyntää terveydenhuollossa. (El-Sappagh ja muut, 2023, s. 11 276)

Generatiivisella tekoälyllä on rajoitettu pääsy lääketieteelliseen tietoon. Se pystyy tunnistamaan kliinisesti merkittävät tekijät, kun ne on erikseen tarjottu, mutta tilanteiden monimutkaistuessa se jättää oleellisia asioita huomioimatta. (Sohail ja muut, 2023, s. 7,

10). Sanmarkin ja Sanmarkin (2024) kirjallisuuskatsauksessa käytettyjen tutkimusten mukaan generatiivinen tekoäly todettiin käyttökelpoiseksi, mutta se suoriutui kokenutta ammattilaista heikommin esimerkiksi hoidon tarpeen arvioinnissa, radiologisten kuvantamistutkimusten valinnassa, tekstin tulkitsemisessa ja kääntämisessä maallikkokielelle, sekä yläraajaortopedisen diagnoosin ja hoitolinjan valinnassa. Heidän analysoimissaan tutkimuksissa ei ilmennyt generatiivisen tekoälyn käyttöön liittyviä merkittäviä riskejä, jotka ehdottomasti estäisivät käytön. Esimerkiksi GPT4-kielimalli ei tehnyt hoidon tarpeen arvioinnissa liian uskaliaita päätöksiä yhtään enempää, verrattaessa käytössä oleviin apuohjelmistoihin. (Sanmark & Sanmark, 2024)

3 Tutkimusmenetelmä

Tutkimusmenetelmänä toimii kirjallisuuskatsaus, joka on valittu tietojärjestelmätieteen kandidaatintutkielman tutkimusmenetelmäksi. Salmisen (2023, s. 4) mukaan kirjallisuuskatsaus on systemaattinen, täsmällinen ja toistettavissa oleva menetelmä, jolla tunnistetaan, arvioidaan ja tiivistetään jo julkaistua tutkimusaineistoa, sekä tehdään niistä johtopäätöksiä. Kirjallisuuskatsaus on kvalitatiivisten ja kvantitatiivisten metodien yhdistelmä eli mixed method. (Salminen, 2023, s. 4). Salminen (2023, s. 3) viittaa Baumeisteriin ja Learyyn sekä Snyderiin, joiden mukaan kirjallisuuskatsauksen tavoitteena on kehittää ja arvioida olemassa olevaa teoriaa, sekä rakentaa uutta. Kirjallisuuskatsaus tarjoaa kattavan kuvan tietystä asiakokonaisuudesta, sekä auttaa tunnistamaan ongelmia. Tieteellinen tutkimus voi olla sirpaloitunutta, ja kumulatiivisuus voi puuttua, jolloin yhteenvetojen tekeminen kirjallisuuskatsausten avulla on perusteltua. Snyder (2019, s. 333) viittaa Tranfieldiin ja muihin, joiden mukaan perinteisen kirjallisuuskatsauksen riskinä on katsauksen pintapuolisuus ja epäsystemaattisuus. Tämä voi johtaa tiedonpuutteeseen, jolloin riskinä on, että tekijät rakentavat tutkimuksen virheellisten olettamusten pohjalta.

Kirjallisuuskatsaus voidaan jaotella monin eri tavoin, kaiken kaikkiaan erilaisia tyyppejä on noin neljäkymmentä. Salminen (2023, s. 7) esittelee kirjassaan kolme perustyyppiä: kuvailevan ja systemaattisen kirjallisuuskatsauksen, sekä meta-analyysin. Tämän tutkielman menetelmänä on kuvaileva, tarkemmin rajattuna narratiivinen kirjallisuuskatsaus. Salmisen (2023, s. 7–8) mukaan se ei rajaa aineiston valintaa tiukasti, ja tutkimuskysymykset ovat väljempiä kuin systemaattisessa katsauksessa. Narratiivinen kirjallisuuskatsaus antaa laajan kuvan käsiteltävästä aiheesta ja sen kehityksestä (Salminen, 2023, s. 7–8). Tätä tukee myös Juntunen ja Lehenkari (2021, s. 336), jotka viittaavat useisiin lähteisiin, joiden mukaan kuvaileva kirjallisuuskatsaus muodostaa kattavan kuvailevan synteesin aiemmista julkaisuista.

Narratiivinen muoto sopii tähän tutkielmaan hyvin, sillä tavoitteena on luoda yleiskuva tekoälyn luotettavuuteen liittyvistä tekijöistä terveydenhuollossa, sen haasteista, ratkaisuista ja kehitysehdotuksista. Tekoäly aihealueena on nopeasti kehittyvä, uutta

tutkimustietoa tulee tarjolle vauhdikkaasti. Tämän vuoksi on hyödyllistä koota synteesi nykytilanteesta. Siitä hyötyvät tekijä henkilökohtaisesti, kanssaopiskelijat sekä mahdollisesti myös terveydenhuolto. Tätä tukee myös (Snyder, 2019, s. 333) perustellessaan, kuinka tiedontuotto kauppatieteellisellä alalla kiihtyy valtavalla vauhdilla, ollen samalla hajanainen ja monitieteinen. Tästä syystä uusimpien tulosten ja teknologioiden suhteen on vaikea pysyä ajan tasalla, sekä arvioida kollektiivista näyttöä tietyllä kauppatieteellisellä tutkimusalueella. Siksi kirjallisuuskatsaus tutkimusmenetelmänä on tärkeämpi kuin koskaan ennen. (Snyder, 2019. s. 333)

Tämän tutkielman aiheenvalinta tehtiin syyskuussa 2024 ja teoriatietoon tutustuttiin loka-joulukuussa 2024. Tutkimusaineisto etsittiin lokakuussa 2024 ja hakua täydennettiin helmikuussa 2025. Tässä työssä käytettävä artikkeliaineisto on etsitty Web of Science -tietokannasta. Hakutermeinä oli käytössä "artificial intelligence" AND "trustworthy AI" AND (healthcare OR medicine). Haku kohdistettiin otsikoihin, ja hakua rajataan englanninkielisiin vuosina 2021–2024 julkaistuihin artikkeleihin, jotka eivät olleet kirjallisuuskatsauksia. Web of Science antoi näillä rajauksilla 30 osumaa lokakuussa 2024. Haku toistettiin 4.2.2025 ottaen myös vuosi 2025 hakuun mukaan. Osumia tuli 34kpl. Uusimmatkin artikkelit otettiin mukaan aineistonvalintaprosessiin.

Artikkelien tiivistelmät lukemalla niiden joukosta löytyi tutkielman tavoitteita tukeva joukko, 15 artikkelia analysoitavaksi (liite 2). Joukosta karsittavaksi päätyivät potilashaastatteluihin pohjautuvat tutkimukset sekä syvästi tiettyyn erityisalaan tai yksityiskohtaan keskittyvät tutkimukset. Mukaan valittujen artikkelien julkaisijat tarkistettiin, jotta kaikki täyttivät vähintään JUFO-tasoluokan 1 tai vastaavan pohjoismaisen tasoluokittelun. Kun valitut viisitoista artikkelia oli luettu, poissulkukriteerit tarkistettiin uudelleen muiden artikkelien osalta. Aineiston analyysissa käytettiin luokittelua ja teemoittelua. Teemoittelu on yksi laadullisen tutkimuksen analyysimenetelmistä ja se on yksi sisällönanalyysin muodoista, siinä aineistosta paikannetaan tutkimuskysymysten kannalta olennaiset aiheet eli teemat (Tuomi & Sarajärvi 2018). Kolmen artikkelin lukemisen jälkeen oli selvää, että artikkeleista nousi selkeät teemat, joiden alle aineistoa alettiin luokitella. Teemoina

nousivat haasteet, teknisemmät mallit, muut niin sanotut pehmeämmät kehitysmenetelmät ja tulevaisuudennäkymät. Lisäksi taulukoitiin se, miten luotettavan tekoälyn eri vaatimukset esiintyivät artikkeleissa. Aineistoa läpikäydessä tehtiin myös kattavat muistiinpanot, joihin lopuksi merkittiin eri värein teemojen mukaiset sisällöt, ja sen pohjalta pystyttiin tuottamaan tulosluvun teksti varmistaen, että kaikki oleellinen tuli mukaan.

4 Luotettava tekoäly terveydenhuollossa

Procterin ja muiden (2023, s. 26-27) mukaan luonnollinen vastuu on ihmisten välisen luottamuksen perusta, ollen tehokkain perusta myös luotettavalle tekoälylle. Vastuullisuus tulee siis upottaa rutiineihin ja toimintoihin, tekoälyjärjestelmän tulee olla vastuullinen terveydenhuoltopolun jokaisessa vaiheessa. Se on kuitenkin monimutkaisen ja laajenevan tietomäärän vuoksi haaste niin tiedon visualisoinnissa, ihmisen ja teknologian välisen vuorovaikutuksen suunnittelussa kuin moniammatillisessa yhteistyössä. Fehr ja muut (2022) mainitsevatkin, että terveydenhuollon tekoälyn hallintomalleja ja tukirakenteita rakennettaessa luottamus ei ole vain teknologian toimimisen varassa, vaan tueksi tarvitaan tietoisia rakenteita ja toimintatapoja. Taulukkoon 1 on koottu kooste siitä, mitä luotettavan tekoälyn vaatimuksia eri tutkimusaineiston artikkelit käsittelivät. Ehdottomasti yleisimmin käsittelyssä oli läpinäkyvyys. Yhteiskunnallinen ja ekologinen hyvinvointi taas näyttäytyi artikkeleissa kaikkein vähiten ja pintapuolisimmin.

Taulukko 1. Luotettavan tekoälyn vaatimusten esiintyminen tutkimusartikkeleissa

Luotettavan tekoälyn vaatimus	Englanniksi	Artikkelin ID
Tulkittavuus	Interpretability	[3], [8], [13]
Tekninen vastustuskyky	Robustness	[2], [4], [7], [9], [14]
Läpinäkyvyys	Transparency	[1], [3], [4], [7], [8], [9], [11], [12], [14], [15]
Selitettävyys	Explainability	[5], [6], [8], [14]
Vastuuvelvollisuus	Accountability	[4], [7], [9], [12], [14]
Vääristymien ja harhojen välttäminen	Avoidance of unfair bias	[4], [5], [9], [12], [14]
Yksityisyys ja datanhallinta	Privacy and data governance	[4], [9], [10], [12], [14]
Oikeudenmukaisuus	Fairness	[4], [9], [10], [12], [14]
Ihmisen toimijuus ja ihmisen suorittama valvonta	Human agency and oversight	[4], [9], [14]
Yhteiskunnallinen ja ekologinen hyvinvointi	Societal and environmental well-being	[4], [9], [14]

Tutkimustulosten käsittely alkaa esiin nousseiden haasteiden läpikäynnillä, jatkuen luotettavuuden arvioinnissa käytettyjen mallien ja muiden menetelmien käsittelyyn.

4.1 Nykytilanteessa tunnistetut haasteet terveydenhuollon tekoälyn luotettavuudessa

Luotettavuuteen haasteet terveydenhuollon tekoälytyökaluissa ovat monitahoisia haasteita, kuten kuviossa 3 ja tässä luvussa esitetään. Bürgerin ja muiden (2024, s. 2) mukaan yksi syy luotettavien tekoälyjärjestelmien epäonnistumiseen on operatiivisen tason puute luottamuksen määrittelyssä. Tämä voi johtaa niin tahattomaan termien väärinkäyttöön, kuin tahallisen väärinkäytön riskiin ja etiikkapesuun alan sidosryhmien toimesta. Eettisestä väärinkäytöstä on huolestuttavia raportteja, esimerkiksi teknologiayritykset keräävät enemmän arkaluontoista potilasdataa kuin julkisesti kertovat ja algoritmit aliarvioivat sairastumisriskiä tai itse sairauksia aliedustetuissa potilasryhmissä, pahentaen rakenteellista syrjintää. (Bürger ja muut, 2024, s. 2).

Zicarin ja muiden (2021b, s. 12) mukaan haasteena on abstraktien periaatteiden soveltaminen yksittäisiin tapauksiin. Tämä edellyttää laajojen käsitteiden rajaamista tulkinnoiksi, jotka ovat hyödyllisiä juuri kyseisessä tapauksessa. Tämä kuitenkin rajaa väistämättä filosofisen, eettisen ja oikeudellisen keskustelun laajuutta ja syvyyttä. Luottamuksen kohteen siirtyessä suorittajasta teknologiapohjaisen tukijärjestelmän tarjoajaan, neuvottelut siitä mitä lääketieteellinen hoito tulisi olla, voivat kietoutua yhä enemmän ulkoisiin kannustinrakenteisiin, kuten voitontavoitteluun. Neuvotteluja voivat myös hajautua ja etääntyä erityisesti haavoittuvassa asemassa olevien potilaiden näkökulmasta (Herzog ja muut, 2024). Bürgerin ja muiden (2024, s. 2) mukaan eettisten periaatteiden ja eettisten käytänteiden välillä on kuilu, joka johtuu aiheen abstraktista luonteesta ja rajallisesta soveltuvuudesta käytännön tekoälyjärjestelmien tutkijoihin ja kehittäjiin. 75–80 % eettisistä ohjeista tarjoaa vain vähän yksityiskohtia ja korkeintaan matalimman tason käytännön näkemyksiä. Esimerkiksi oikeudenmukaisuuden arviointiin on olemassa yli 20 erilaista mittaria, joista osa on ristiriitaisia.

Kuvio 3. Tunnistetut haasteet terveydenhuollon tekoälyn luotettavuudessa

Haitallisuuden riskin arvioinnissa tulisi Zicarin ja muiden (2021a, s. 6, 9) mukaan selvittää mikä määrä vääriä positiivisia ja negatiivisia on hyväksyttävissä, onko olemassa standardoitua tapaa määritellä kulut eri sidosryhmien näkökulmasta ja kuinka määritellään, onko tekoälyjärjestelmä haitallinen. Esimerkiksi exAID-malli (Explainable AI in Dermatology) on koulutettu rajallisella määrällä julkisesti saatavilla olevia ihokuvia, joissa esiintyy laatuvariaatioita ja artefakteja, eivätkä ne välttämättä edusta potilaan taustoja. Tummemmat ihosävyt puuttuvat kokonaan koulutusdatasta, sillä melanooma on 20–30 kertaa harvinaisempi tummemmilla ihosävyillä, mutta kuolleisuus on heillä kuitenkin korkeampi kuin valkoihoisilla. (Zicari ja muut, 2021a, s. 6, 9)

Läpinäkyvyyden ja mustan laatikon ongelmat ovat keskeisiä haasteita. Fehr ja muut (2024, s. 7–8) tutkivat neljäntoista CE-sertifioidun tekoälypohjaisen radiologian tuotteen

julkista dokumentaatiota todeten sen olevan turvallisuuden ja riskien läpinäkyvyyden kannalta puutteellista. Dokumentaatiopuutoksia oli erityisesti koulutusdatan, eettisten huomioiden, käyttöönoton rajoitusten suhteen. Karim ja muut (2022, s. 54 407) nostavat esiin MCAE-mallinsa (Multimodal Convolutional AutoEncoder) rajoitteeksi syväoppimismallien läpinäkymättömyyden ja sitä kautta heikon jäljitettävyyden. Zicari ja muut (2021b, s. 12–13) korostavat, että läpinäkyvyys on haaste kaikille sidosryhmille. Esimerkiksi hätäkeskuspäivystäjillä olisi oltava tietoa siitä, mitä arvoja on käytetty tekoälyn suosituksen tekemiseen, jotta he voivat arvioida onko tekoälyn aktivoima hälytys sydänpysähdyksestä pätevä vai ei, ja voivatko he itse kyseenalaistaa päätökset. Tietosuojalainsäädäntö GDPR voidaan ohittaa, jos tilanne on yksilölle elintärkeä tai kohde ei pysty antamaan suostumusta, mutta laillista arviointia tarvitaan tilanteessa, jossa kohteella ei ole tietoa tekoälyjärjestelmästä eikä myöskään mahdollisuutta suostumuksen antamiseen. Päivystäjien toimijuus ja itsenäinen päätöksenteko näyttivät myös vähenevän järjestelmän myötä, mikä madalsi sekä päivystäjien sitoutumista että toimijuuden tunnetta. (Zicari ja muut, 2021b, s. 12, 14)

Fehr ja muut (2022, s. 1) toteavat, että liiketoimintarajoitteet ja ulkoisisten aineistojen rajoitteet estävät läpinäkyvyyttä kaupallisissa tekoälytyökaluissa, mikä heikentää niiden luotettavuutta ja eettisten ohjeiden noudattamista. Toimijat perustelivat rajoitteita sillä, että käytetyn datan raportointi uhkaa kilpailuetua, ja että tietoja voitiin paljastaa vain sääntelyhyväksynnän saamiseksi. Tutkijat pohtivatkin, onko luotettavuuden ja läpinäkyvyyden ulkoinen auditointi mahdollista vasta markkinoille tulon jälkeen, kun patentit on turvattu. Zicari ja muut, (2021a, s. 21) kuitenkin huomauttavat läpinäkyvyyden auttavan rahoittajia, valvontaviranomaisia ja johtoryhmiä selittämään päätöksensä.

Lin ja muiden (2024, s. 4584) mukaan vastatodelliset selitykset tarjoavat tavan tutkia ”entä jos” -skenaarioita. Kun tekoäly tarjoaa diagnoosia, vastatodellinen selitys kertoo mitä tietojen olisi pitänyt olla, jotta diagnoosi olisi muuttunut. Nykyiset tavat luoda vastatodellisia selityksiä eivät ole luotettavia, sillä ne voivat tuottaa epärealistisia esimerkkejä, joita ei oikeasti voisi esiintyä, tekoäly voidaan huijaa tekemään virheellisiä

päätöksiä eivätkä ne tarjoa epävarmuusarviota eli ei kerrota kuinka luotettava selitys on. (Li ja muut, 2024, s. 4584). Zicari ja muut (2021a, s. 17) huomauttavat, että moderni lääketiede ei perustu vain tieteeseen ja kliinisiin suosituksiin, siihen vaikuttavat myös perinteet, kulttuuri ja erilaiset tulkinnat tutkimustiedosta. Kansallisellakin tasolla voi olla erilaisia suosituksia ja strategioita. Kehitystiimien tulee olla tästä tietoisia, muutoin tekoälytyökalu saattaa olla käytettävissä vain rajoitetulla alueella tai tietyn kulttuurin tai uskonnon edustajilla.

Zicarin ja muiden (2021a, s. 12) mukaan aiemmissa tutkimuksissa on oletettu kaikkien harhojen tekevän tekoälytyökalusta automaattisesti eettisesti kestämättömän. He kuitenkin huomauttavat, että seurausetiikan näkökulmasta se on ongelmallista vasta, kun harhan aiheuttama haitta ylittää tekoälytyökalun tuoman hyödyn. Zicari ja muut (2021b, s. 12–13) korostavat, että harha ja reiluus ovat alakohtaisia ja ne tulee huomioida eri tasoilla terveydenhuollon toimijoiden näkökulmasta aina koneoppimismallin tasolle asti. Tanskassa testattu tekoälyjärjestelmä sydänpysähdyksen tunnistamiseen hätäpuhelussa esimerkiksi antoi enemmän vääriä negatiivisia, jos soittaja oli eri huoneessa kuin potilas tai puhui vahvalla murteella tai muuta kuin tanskaa. Testidata myös painottui vanhempiin miehiin, mutta he ovat myös yleisin potilasryhmä, joten Zicari ja muut pohtivatkin, onko tässä tilanteessa kyseessä harha.

Zicari ja muut (2021b, s. 13) nostavat esiin hätäkeskuspäivystäjän oikeudellisen vastuun tekoälyjärjestelmää käytettäessä, päätöksenteon perustelu sekä ihmisen ja tekoälyjärjestelmän välinen vuorovaikutus ovat oleellisia. Riskinä on osaamisen heikkeneminen ja teknologinen delegointi, koska ei haluta vastuuseen hälytyksen huomiotta jättämisestä tai hylkäämisestä. Vain viidesosa tekoälyjärjestelmän hälytyksistä on todellisia sydänpysähdyksiä. Näin alhaisessa herkkyydessä on riski hälytysväsymykseen, eli että päivystäjät jättäisivät huomiotta todelliset hälytyksetkin. Täysin autonominen järjestelmä saattaisikin olla turvallisempi kuin sellainen, jossa on liian monta inhimillistä päätöksentekovaihetta, sillä varsinkin hälytysten huomiotta jättäminen vaatisi koneoppimismallin toiminnan ymmärrystä hätäkeskuspäivystäjältä. Zicari ja muut (2021b, s. 14) jatkavat

vastuuvelvollisuudesta. Terveydenhuollossa, erityisesti elämän ja kuoleman kysymyksissä mahdollinen haitta on huomattava taloudellisesti ja muutenkin. ALTAI:ssa vastuuvelvollisuus on eniten ei-oikeudellista. Algoritmien tuntemattomuuden takia voidaan tehdä vain geneerisiä ohjeita, ja oikeudellisten ja ei-oikeudellisten välillä on väistämättä vuorovaikutusta (Zicari ja muut, 2021b, s. 14). Alalta puuttuukin välineet ja kannustimet muuttaa korkeatasoiset eettiset periaatteet todennettaviksi ja toimiviksi kriteereiksi (Pourzolfaghar ja muut, 2023, s. 125).

Yleisesti tekoälyn luotettavuus nähdään vain ihmisen ja tietokoneen vuorovaikutuksen haasteena, mikä jättää organisaatiotason vastuun vähemmälle huomiolle, vaikka se kuitenkin vaikuttaa siihen, miten ihmisten luottamus muodostuu sosioteknisessä ympäristössä (Procter ja muut, 2023, s. 1). Procter ja muut (2023, s. 28) huomauttavat, että myös organisaatioiden väliset tekijät on otettava huomioon. He viittaavat Selleniin ja Harpeniin, joiden mukaan uuden teknologian käyttöönotto organisaatioympäristössä edellyttää käytäntöjen ja teknologian uudelleenkonfigurointia, jotta teknologiaa voidaan käyttää ja mukauttaa tarpeen mukaan. Procterin ja muiden (2023, s. 23–25) mukaan luotettavassa ja selittävässä tekoälyssä erotetaan globaalit ja paikalliset selitykset. Organisaatiokontekstissa vastuullisuus taas jaetaan muodolliseen, tilannesidonnaiseen ja luonnolliseen. He esittävät viisi skenaariota: 1) globaali selittävä tekoäly ja muodollinen vastuullisuus tulosten validointiin ja tarkkuuteen liittyen, 2) globaali vastuu voi olla riittämätön yksittäistapauksissa, 3) hoidon tarpeen arviointiohjelmiston tapaustutkimuksessa radiologit luottivat vain muistiinsa, 4) moniammatillisten kokouksien rooli organisaatiovastuun solmukohtana (otollinen paikka tekoälyjärjestelmälle), 5) uusien teknologioiden käyttöönoton haasteet kontekstisidonnaisissa kysymyksissä. (Procter ja muut, 2023, s. 23–25)

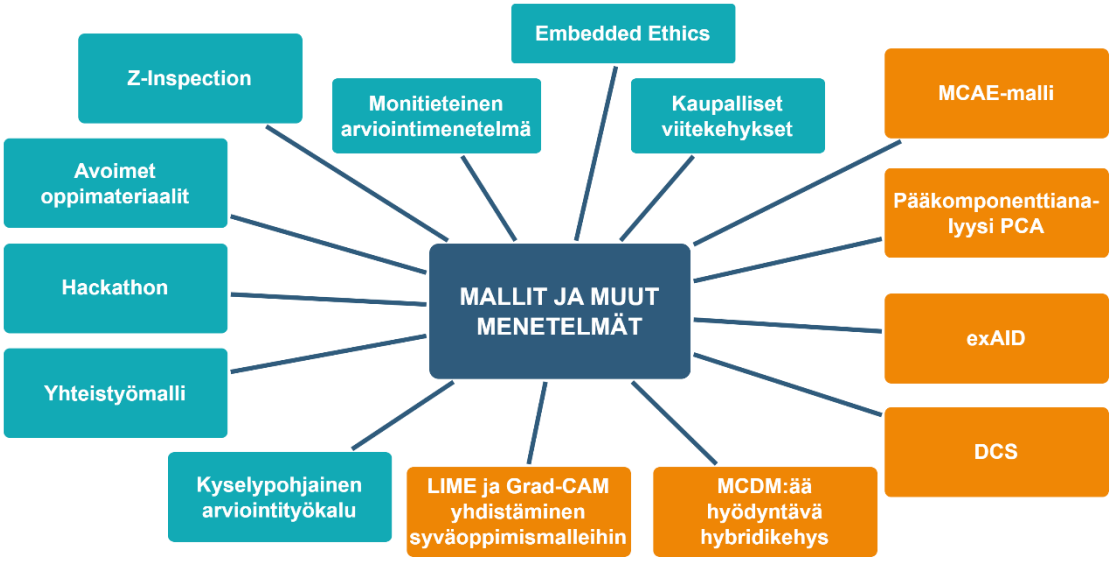
Herzog ja muut (2024) nostavat esiin kuilun periaatteiden ja käytännön välissä. Bürgerin ja muiden (2024, s. 1) mukaan kliinisessä todellisuudessa ja tekoälytyökalujen saatavuudessa on vahva ristiriita, eikä yhtäkään työkalua ole sisällytetty vakiintuneesti kliinisiin lääketieteellisiin ohjeisiin tai käytännön normeihin. Silloinkin kun työkaluja olisi saatavilla,

ne eivät vastaa eettisiä vaatimuksia. Bürgerin ja muiden (2024, s. 2) mukaan kartoittavia tutkimuksia on paljon, mutta niiden tuloksia ei ole validoitu konfirmatorisilla tutkimuksilla, eli hypoteeseja ei ole testattu empiirisesti, jotta löydökset olisi turvallista siirtää kliiniseen käyttöön. Lääketieteellisissä tuotteissa tätä kehityskaarta kutsutaan translaatioksi, ja tutkimuksen ja kliinisen käytännön välistä kuilua translaatiovajeeksi. Zicari ja muut (2021a, s. 3) kertovat, että tekoälyjärjestelmät ovat tilastollisesti saavuttaneet ihmistason suorituskyvyn ihosyövän kuvatunnistuksessa, ja he viittaavat edelleen Brinkeriin, jonka tutkimuksen mukaan CAD (Computer-Aided Diagnosis) on saavuttanut jo paremman herkkyyden ja tarkkuuden kuin hyvin koulutetut ihotautilääkärit. Heikon selitettävyyden vuoksi näitä järjestelmiä on kuitenkin kliinisessä käytössä hyvin rajatusti. Alsalem ja muut (2024, s. 13) tuovat esiin ettei DCS (Deep Conformal Supervision) huomioi asiantuntijoiden näkökulmien tai osaamisalueiden suhteellista painoarvoa, ja se voi heikentää tulevia sovelluksia. Fehrin ja muiden (2022, s. 8) mukaan huonosti raportoituja lääketieteellisiä tekoälytyökaluja on yleisesti käytössä, monista saatavilla olevista ohjeista huolimatta. Pourzolfaghar ja muut (2023, s. 125) mainitsevat, että tekoälyjärjestelmiä on viime vuosina kehitetty tavoilla, jotka eivät ole linjassa sen luojien arvojen kanssa, mikä herättää huolta niiden käytön vaikutuksesta ja hyödyllisyydestä.

Tekoälyn käyttöön liittyy myös metaeettisiä riskejä, eli metaeettiset edellytykset kuten myötätunto, empatia, solidaarisuus ja epäoikeudenmukaisuuden tunnistaminen voivat muuttua perustavanlaatuisesti tekoälyä käytettäessä. Riskit muodostuvat ihmisten välisen vuorovaikutuksen, inhimillisen haavoittuvuuden ja pohdinnan puutteesta. Kun osa viestinnästä siirretään algoritmeille, tehokkaamman terveydenhuollon lisäksi se voi heikentää ammattilaisten sitoutumista potilaiden haasteisiin (Bærøe & Gundersen, 2019, s. 12–13). Zicari ja muut (2021a, s. 12) mainitsevat myös, että riskien ja hyötyjen avaaminen viestinnän keinoin voi lisätä luottamusta.

4.2 Terveydenhuollossa käytetyn tekoälyn luotettavuuden parantamiseksi löydetyt ratkaisut

Tutkimusaineisto tarjosi lukuisia malleja ja muita menetelmiä käytettäväksi tekoälyn luotettavuuden arvioinnissa. Aineistosta nousi selkeästi teknisempiä malleja (kuviossa 4 oranssit), sekä niin sanottuja pehmeämpiä menetelmiä (kuviossa 4 turkoosit). Nämä ratkaisut käydään läpi tässä alaluvussa yksitellen, aloittaen pehmeämmistä menetelmistä. Luvun lopussa aihe kootaan Herzogin ja muiden (2024) sosioteknisen ekosysteemiajattelun avulla.



Kuvio 4. Terveydenhuollon tekoälyn luotettavuuden arviointiin kehitettyjä malleja ja muita menetelmiä

Useat tutkimuksista nostavat esiin yhteistyön ja moniammatillisen työskentelyn merkityksen (Pourzolfaghar ja muut, 2023; Zicari, 2021a; Procter ja muut, 2023; Gundersen & Bærøe, 2022; Bürger ja muut, 2024; Herzog ja muut, 2024). Monitieteinen arviointimenetelmä voi auttaa tunnistamaan huolenaiheet useilla eri aloilla jo kehityksen varhaisessa vaiheessa (Zicari ja muut, 2021a, s. 16). Pourzolfaghar ja muut (2023, s. 113) totesivat Hackathonin tarjoavan toimivan tavan tuottaa suosituksia tekoälyratkaisuille terveydenhuoltoon, ja osoittavan avoimien oppimateriaalien tehokkuuden ja

soveltuvuuden. Tutkimuksessa Hackathoniin osallistuivat opiskelijat ja tiedekunnan jäsenet tiimeissä, arvioiden miten CHAPE :a eli käyttäjän kanssa luonnollisella kielellä terveydestä ja hyvinvoinnista keskustelevaa agenttia voidaan muokata ja sopeuttaa vastaamaan luotettavan tekoälyn seitsemää eettistä vaatimusta.

Procter ja muut (2023, s. 9) viittaavat Hartswoodiin ja muihin viitatessaan varhaisiin tutkimuksiin, joissa tekoälyn käyttö rintasyövän tunnistamisessa korostaa kliinikoiden yhteistyöhön perustuvaa ammatillista näkemystä, kun he tulkitsevat teknologian tuottamia tuloksia. Myöhemmät tutkimukset ovat vahvistaneet tätä löydöstä (Procter ja muut, 2023, s. 9). Gundersen & Bærøe (2022, s. 1-3) puolestaan korostavat lääkäreiden, tekoälysuunnittelijoiden ja muiden sidosryhmien yhteistyötä tekoälyn eettisessä hyväksyttävyydessä, erityisesti lääketieteellisen yhteispäätöksenteon (shared decision-making) näkökulmasta. Tutkijoiden mukaan yhteistyömalli (collaborative model) on lupaavin vaihtoehto suurimmalle osalle tekoälyteknologioista. Sen mukaan tekoälysuunnittelijoiden ja lääkäreiden välinen keskinäinen sitoutuminen ja kommunikaatio ovat välttämättömiä algoritmien sovittamiseksi lääketieteeseen. Mallin tavoitteena onkin kuroa umpeen kuilu näiden toimijoiden välillä. (Gundersen & Bærøe, 2022, s. 10)

Gundersen & Bærøe (2022, s. 11) esittelevät kolme tapaa toteuttaa yhteistyötä: lääkäreiden osallistuminen tekoälyn suunnitteluun jo alkuvaiheesta lähtien, tekoälykehittäjien ja lääkäreiden yhteistyö tekoälyn tuottamien tulosten ymmärtämiseksi ja tulkitsemiseksi kliinisessä käytössä, sekä vaikutusten arviointi kliinisessä käytössä. Molemminpuolinen osallistuminen parantaa sekä tekoälykehittäjien lääketieteellistä lukutaitoa että lääkäreiden tekoälylukutaitoa. Esimerkiksi jotkut algoritmit tarjoavat liian paljon vääriä positiivisia, jolloin lääkärien palaute suunnitteluprosessissa voisi vähentää algoritmin ”innokkuutta” sairauden merkkien havaitsemisessa. Lääkäreillä ja bioeetikoilla tulee olla selkeät väylät esittää kritiikkiä, huolia ja parannusehdotuksia suunnittelijoille. Julkisen keskustelun malli puolestaan sisältää sidosryhminä myös päättäjät ja kansalaiset, ja sitä tarvitaan silloin kun teknologia on muuttamassa perustavanlaatuisesti eettisen yhteispäätöksen teon edellytyksiä. (Gundersen & Bærøe, 2022, s. 12)

Bürger ja muut (2024, s. 4) korostavat sosioteknisen näkökulman tärkeyttä, eli sitä että sekä tekoälyjärjestelmän että ihmissidosryhmien näkökulmat huomioidaan. Eettisten periaatteiden käytäntöön siirtämiseen on kehitetty useita viitekehyksiä, esimerkiksi Z-Inspection® ja Embedded Ethics, joka keskittyy eettisten periaatteiden integrointiin koko teknologian kehitysprosessissa kannustaen esimerkiksi filosofien osallistamiseen tekoälyohjelmistojen kehitystiimeissä. Näiden avoimesti saatavilla olevien viitekehysten lisäksi tarjolla on myös kaupallisia menetelmiä, kuten Digital Catapult.

Z-Inspection® on kokonaisvaltainen arviointimenetelmä, jonka mukaan eettisistä kysymyksistä on keskusteltava laatimalla sosioteknisiä skenaarioita. Se koostuu käyttöönottovaiheesta, arviointivaiheesta ja ratkaisuvaiheesta. Käyttöönottovaiheessa valitaan asiantuntijajoukko sidonnaisuudet huomioiden, arviointivaiheessa luodaan ja analysoidaan sosiotekniset skenaariot, tunnistetaan eettiset ongelmat, kartoitetaan ne ALTAI:n puitteissa ja todennetaan vaatimukset. Ratkaisuvaiheessa ratkaistaan mahdollisuuksien mukaan esiin nousseet jännitteet, ja annetaan sidosryhmille suositukset. (Zicari ja muut, 2021a, s. 3–4). Z-Inspection on yhteissuunnitelmallinen (co-design) lähestymistapa, jonka avulla on pystytty tunnistamaan ongelmia, joita ei ollut mahdollista tunnistaa perinteisen suunnittelutyöskentelyn kautta. Se tarjoaa yhteistyön suunnittelijoiden ja pääsidosryhmien välillä, jotta tunnistetaan erilaiset käyttötavat, mitkä ovat niiden edut ja haitat ja mitkä käyttötavoista on ensisijaista käyttöä. (Zicari ja muut, 2021a, s. 5).

Zicari ja muut (2021b, s. 2–3) käyttivät Z-Inspectionia sydänpysähdyksiä tanskalaisista hätäpuheluista tunnistavan tekoälyjärjestelmän arviointiin. He totesivat, että sitä voi käyttää riskien arviointiin, eettisten jännitteiden tunnistamiseen, prosessien laadunparannukseen sekä läpinäkyvyyden lisäämiseen eturistiriitojen suhteen. Se myös lisää järjestelmän ymmärrettävyyttä, joka parantaa viestinnän laatua kaikille sidosryhmille. Arviointiprosessista tiedottaminen voi auttaa luottamuksen vahvistamisessa, kun läpinäkyvyys paranee myös ei-ammattilaisille (Zicari ja muut, 2021b, s. 21). He muodostivat viisi suositusta tapaustutkimuksen pohjalta: 1) päivystäjien tarve ymmärtää mallin ennusteita, 2) testidatan tietoinen poiminta tai heuristiikka, joka ilmoittaa milloin mallia voi ja

ei voi käyttää, ja vajaan luotettavuuden osalta ilmoitus siitä, 3) sidosryhmien osallistaminen, 4) tarvitaan lisätutkimusta siitä, mihin hätäkeskuksen kyselyprotokolla mahdollisesti vaikuttaa ja 5) järjestelmällä on tarve CE-merkinnälle. (Zicari ja muut, 2021b, s. 19)

Fehr ja muut (2022, s. 1, 9) kehittivät kyselypohjaisen arviointityökalun, jolla voidaan kvantifioida lääketieteellisten tekoälysovellusten läpinäkyvyyttä. Arviointityökalu koostuu 78 kysymyksestä, joiden avulla lasketaan läpinäkyvyyden ja luotettavuuden aste prosentteina. Kyselylomakkeella raportoidaan aiottu käyttö, koulutus- ja validointidatat sekä prosessit, eettiset näkökulmat ja käyttöönoton suositukset. Arviointityökalu pilotoitiin kolmessa käyttötapauksessa, joille tehtiin ulkoinen auditointi. Tarve laajemmalle soveltamiselle havaittiin liiketoimintarajoitteiden suojaamiseksi terveydenhuollon tekoälyjärjestelmissä, jotta ulkoiset auditoinnit olisivat mahdollisia. (Fehr ja muut, 2022, s. 9)

Li ja muut (2024, s. 4584, 4597) kehittivät uuden menetelmän, joka soveltaa pääkomponenttianalyysi PCA:ta differentiaaliseen generatiiviseen malliin. Tällä uudella metodilla vastatodelliset selitykset ovat vastustuskykyisiä virheille ja häiriöille, sekä sisältävät epävarmuusanalyysin. Metodi testattiin keuhkoröntgenkuvilla ja kasvokuva-aineistolla, ja se suoriutui muita metodeja paremmin. Li ja muut (2024, s. 4598) arvioivatkin sen voivan olla tärkeä osa selittävää tekoälyä, sillä se mahdollistaa monimutkaisen tiedon tiivistämisen tärkeimpiin piirteisiin parantaen luotettavuutta, tarkkuutta ja läpinäkyvyyttä.

Karim ja muut (2022, s. 54407) puolestaan kehittivät MCAE-mallin (Multimodal Convolutional AutoEncoder), joka perustuu latentin edustuksen yhdistämiseen (latent representation concatenation) konvoluutioautokoodereihin (convolutional autoencoders). Sen tarkkuus on 96,25 % syöpäalttiuden ennustamisessa, sillä se käyttää multiomiikkadataa ja on varautunut vihamielisiin hyökkäyksiin. Malli oppii monimuotoisia piirre-edustuksia, joiden avulla se luokittelee potilasryhmät eri syöpätyyppeihin. Se koulutettiin erilaisilla vihamielisten hyökkäysten skenaarioilla käyttäen ennaltaehkäiseviä ja korvaavia toimenpiteitä, kuten mallin uudelleenkoulutusta haavoittuvuuden tunnistamiseksi ja vähentämiseksi, sekä epäilyttävien tietojen suodatusta. Näin ennusteet pysyvät vakaina

pienistä syötteeseen kohdistuvista muutoksista huolimatta, ja malli toimii sekä proaktiivisesti että reaktiivisesti. (Karim ja muut, 2022, s. 54388–54389, 54407). MCAE-mallin rajoituksia ovat rajallinen merkitty data, syväoppimismallien läpinäkymättömyys ja siitä johtuva mallin heikko jäljitettävyys, sekä monimodaalisten edustusten haasteet, eli kun eri tietotyyppejä yhdistetty, on vaikeampi ymmärtää yksittäisten tekijöiden vaikutuksia. (Karim ja muut, 2022, s. 54407)

Zicari ja muut (2021a, s. 3) esittelivät exAID-prototyypin (Explainable AI in Dermatology), joka yhdistää ihokasvainten luokitteluun suunnitellut korkean suoritustason hermoverkot käsitepohjaisiin selitystekniikoihin, tällöin diagnoosiehdotukset ja selitykset vastaavat asiantuntijoiden hyväksymiä diagnoosikriteerejä. Seuraava askel on käyttää altistusanalyysiä loppukäyttäjän haavoittuvuutta vähentämään, sekä osallistaa potilaita suunnitteluprosessin kaikkiin vaiheisiin, jotta tieto on myös heille ymmärrettävässä muodossa (Zicari ja muut, 2021a, s. 13).

Yksi luotettavan tekoälyn peruspilareista on epävarmuuden kvantifiointi, johon on käytetty konformaalista ennustamista (CP) vakiintuneena menetelmänä. Uusi sekamenetelmäksi tulkittava syvä konformaalinen valvonta eli DCS (Deep Conformal Supervision) integroi CP:n valvontaan. Se ei siis arvioi pelkkää epäyhtenäisyyspisteiden (non-confirmity score) CP:tä, vaan tekee arviointia myös syväoppimismallin välikerroksissa ennen lopullista ennustetta. Se käyttää syvempiä tietorakenteita yhdistelemällä tietoja niin, että epävarmuuden arviointi on tarkempaa, parantaen selitettävyyttä. DCS testattiin kahdella julkisella lääketieteen kuvantamisen aineistolla, keuhkokuumeen röntgenkuvadatasetillä ja esikäsitellyllä aivoverenvuodon datasetillä. Tällaiset integroidut lähestymistavat ovat tulevaisuutta, sillä kliinisissä terveydenhuollon sovelluksissa ne vähentävät kattavuusvirheitä eli vääriä negatiivisia seulontatuloksia, parantavat luottamusta kriittisiin diagnooseihin, optimoivat resurssien jakamista ja tarjoavat merkittäviä haittavaikutuksia sisältäviin hoitoihin hyötyä riskiluokittelukehyksen avulla. (Vahdani & Faghani, 2024)

Alsalem ja muut (2024, s. 1, 12-13) esittelevät hybridikehyksen, joka hyödyntää monikriteeristä päätöksentekoa (MCDM, Multi-criteria decision-making) epävarmassa ympäristössä. Se kehitettiin rakentamalla uusi päätösmatriisi, johon integroitiin kaksi menetelmää (q-ROF2TL-FWZIC ja q-ROF2TL-CODAS), joita voidaan käyttää luotettavan tekoälyn arviointikriteerien painotusten määrittämiseen ja sovellusten vertailuun. Kehyksen pätevyyttä arvioitiin systemaattisen rankingin ja herkkyysanalyysin avulla. Se edistää syrjimättömien ja puolueettomien sovellusten kehittämistä huomioiden turvallisuuden, yksityisyyden ja säädöstenmukaisuuden, vaikuttaen erityisesti läpinäkyvyyteen, luottamukseen, eettisiin näkökohtiin ja vastuullisuuteen. (Alsalem ja muut, 2024, s. 13)

Garg ja muut (2025) yhdistivät selittävät tekoälymenetelmät LIME (Local Interpretable Model-Agnostic Explanations) ja Grad-CAM (Gradient-weighted Class Activation Mapping) viiteen syväoppimismalliin (CNN, XceptionNet, EfficientNet, VFF ja ResNet) lääketieteellisen kuvien luokittelutehtävissä. LIME merkitsee kuviin ne alueet, joilla on ollut isoin merkitys mallin päätöksenteossa ja Grad-CAM tuottaa lämpökartan korostaen siten merkityksellisimpiä alueita kuvasta. Tämä kaksiosainen lähestymistapa parantaa päätösten läpinäkyvyyttä ja tulkittavuutta, vahvistaen siten luottamusta. Jatkossa mallia on tarkoitus laajentaa muihin selittävän tekoälyn tekniikoihin, sekä laajempiin ja monipuolisempiin kuva-aineistoihin. (Garg ja muut, 2025)

Mittal ja muut (2024, s. 937-938) tutkivat kuvantamisen tietojoukkojen vastuullisuutta yksityisyyden, oikeudenmukaisuuden ja säännösten noudattamisen eli datan suojauksen näkökulmasta. He antoivat neljä suositusta tietojoukkojen vastuullisuuteen: 1) ihmisiä koskevat aineistot tulisi hyväksyä arviointilautakunnissa, 2) GDPR:n noudattaminen eli mahdollisuus tietojen poistoon tai korjaamiseen tulee olla vakiokäytäntö, 3) aineiston kattavuus, väestöpohja ja herkkien attribuuttien jakaumat tulee toteuttaa yksityisyyttä suojaavalla tavalla, 4) kattava dataseloste eli tavoitteiden, aiotun käytön, rahoittajan, yksilöiden tai kuvien demografisen jakauman, lisensointitietojen ja aineiston rajoitusten ilmoittaminen. (Mittal ja muut, 2024, s. 943)

Herzog ja muut (2024) tarkastelivat luotettavaa tekoälyä ekosysteemien näkökulmasta, eli yksittäistä teknologiaa laajemmasta sosioteknisestä perspektiivistä. Tapaustutkimuksen kohteena oli saksalainen Ki-Med ekosysteemi, jonka tarkoituksena on kääntää tekoälytutkimusta lääketieteellisiksi sovelluksiksi. Tutkijat tunnistivat, että siitä puuttui selkeä arvolupaus ja arvot oli johdettu ylhäältä alaspäin, joten inklusiivinen prosessi yhteisen eettisen vision määrittämiseksi ja säännölliseksi päivittämiseksi, sekä eettinen lautakunta ja ulkoiset auditoinnit ovat tarpeen. Herzogin ja muiden (2024) mukaan ekosysteemin luotettavuus pohjaa kahteen näkökulmaan: rationaaliseen valintaan eli että luottamus nähdään järkevänä, kun palveluntarjoajat tuottavat palveluita luotettavasti ja odotusten mukaisesti, sekä motiivien jakamiseen eli kun ihminen luopuu osittain kontrollista siirtäen sitä esimerkiksi tekoälylle tai sen kehittäjälle. Alustan osia ekosysteemissä tulisikin tarkastella synergisinä toimijoina, jotka tekevät tiivistä yhteistyötä yritysten kanssa sekä jakavat yhteisiä tiedollisia resursseja. Vastuullisesta innovaatioalustasta voi olla apua tutkimukseen, operatiiviseen toteutukseen ja strategiseen suunnitteluun. Herzog ja muut (2024) luettelivat ratkaisuja tekoälyn luotettavuuden parantamiseen: 1) tarvitaan hallintarakenteita ja raportointiprosesseja, jotka heijastavat ekosysteemin monimutkaisuutta eli eri tasot paikallisesta globaaliin on liitettävä asianmukaisesti toisiinsa, 2) sidosryhmien ja ekosysteemien jäsenten tuetusti mukana oleminen tuottavassa vuorovaikutuksessa, 3) yksittäisten alajärjestelmien ja niiden välisen yhteistyön ja kilpailun tasapainottaminen, 4) asianmukainen raportointikulttuuri, 5) eettiset huolenaiheet tulee tunnistaa ja käsitellä, eli eettisten vastuiden jakamisen vahvistaminen tuomalla alustan osat osaksi ekosysteemiä.

5 Johtopäätökset ja pohdinta

Tämän tutkielman tulokset tarjoavat arvokkaan tilannekatsauksen terveydenhuollon tekoälyjärjestelmien luotettavuuden haasteisiin ja nykytilaan. Tulokset korostavat tarvetta tekoälyjärjestelmien varhaiseen arviointiin, moniammatilliseen kehitysyhteistyöhön, auditointeihin sekä kehitettyjen teorioiden ja mallien empiiriseen tutkimukseen. Tutkimustulokset korostivat teorialuvussakin esiin tuotua haastetta siitä, että mallit on rakennettu reaalimaailman asetuksista eristettynä laboratorioissa, esimerkiksi ilman ison sähköisen potilastietorekisterin ekosysteemiä (El-Sappagh ja muut, 2023, s. 11 151). Tutkimustulosten valossa sosioteknisen näkökulman ja ekosysteemiajattelun implementointi terveydenhuollon tekoälyn suunnitteluun ja kehittämisprosessiin on kriittinen osa luotettavuuden parantamista. Euroopan unionin tuoreen tekoälyasetuksen tuoma säätely voi tarjota Pourzolfagharin ja muiden (2023, s. 123) peräänkuuluttamia kannustimia eettisten periaatteiden muuttamisessa todennettaviksi ja toimiviksi kriteereiksi.

Sisäisiä ja ulkoisia auditointeja järjestelmän kehityksen eri vaiheissa suositellaan yleisesti (Fehr ja muut, 2024; Procter ja muut 2023; Fehr ja muut 2022; Herzog ja muut 2024), sillä esimerkiksi terveydenhuollon ammattilaisten päätöksentekoprosessi tai tekoälyjärjestelmän suorituskyky voivat muuttua (Procter ja muut, 2023, s. 25). Tutkijoilla ja ulkopuolisilla auditoijilla tulisi olla pääsy koulutusdataan (Fehr ja muut, 2024, s. 8) ja ulkoisiin auditointeihin tulisi kannustaa jo ennen markkinoille tuloa (Gundersen & Bærøe, 2022, s. 9). Sekä Euroopan komissio että Yhdysvaltain lääketieteellisen informatiikan yhdistys ovat nostaneet esiin huolen läpinäkyvyydestä, käytetyistä standardeista ja koulutuksesta sekä järjestelmien seurannasta. (Sanz ja muut, 2019, s. 7).

Pakollisista läpinäkyvyysvaatimuksista tutkijoilla oli keskenään ristiriitaa. Fehr ja muut (2024, s. 8) suosittelevat niitä osaksi markkinoille saattamista edeltävää lupakäsittelyä. Zicari ja muut (2021b, s. 19) taas toteavat, että terveydenhuollon tekoälyn arvioinnissa ja sääntelyssä tarvitaan joustavuutta, eikä tiettyjen vaatimusten vakiinnuttaminen välttämättä ole suositeltu toimintatapa. He kuitenkin vaativat julkista ja sisällöllistä läpinäkyvyyttä erityisesti lääketieteellisistä tekoälytuotteista kaikille sidosryhmille. Fehrin ja

muiden (2022, s. 10) mukaan poliittista säätelyä ja läpinäkyvyyden vähimmäisvaatimuksia lääketieteellisille tekoälysovelluksille tarvitaan, sillä mitkään työkalut yksinään eivät voi taata luotettavuutta ja läpinäkyvyyttä. Fehr ja muut (2024, s. 8–9) suosittelevat läpinäkyvyysvaatimusten määrittämisessä osallistavaa prosessia, jossa tunnistetaan ja neuvotellaan eri sidosryhmien, kuten potilaiden, terveydenhuollon tarjoajien, kehittäjien, tutkijoiden ja sääntelijöiden, intressejä.

Useampi tutkimus korostikin eri sidosryhmien välistä yhteistyötä. Zicarin ja muiden (2021a, s. 16–17) mukaan pelkät tekniset tiimit ovat epäsopivia ymmärtämään kliinisen päätöksenteon nyansseja, ja hyödyllistä onkin hakea näkökulmia suunnitteluprosessin kaikissa vaiheissa kliinistä työtä tekeviltä lääkäreiltä, lääketieteellisestä tutkimusdatasta, kansanterveydestä ja potilailta (Zicari ja muut, 2021a, s. 6). Zicari ja muut (2021b, s. 22) suosittelevat ALTAI:n muokkaamista kontekstisidonnaiseksi moniammatillisen asiantuntijaryhmän avulla. Laajaa ja kattavaa asiantuntijuutta tarvitaan kaikilla terveydenhuollon tekoälyn osa-alueilla. Gundersen & Bærøe (2022, s. 12) suosittelevat tekoälypohjaisen päätöksenteon kouluttamista terveydenhuollon koulutusohjelmissa, ja kurssien kehittäminen yhteistyössä sidosryhmien kanssa.

Hyvin sovitettu ja hyökkäyksenkestävä malli voi tarjota johdonmukaisia ja luotettavia diagnooseja (Karim ja muut, 2022, s. 54407), tarjoten yhden ratkaisun El-Sappaghin ja muiden (2023, s. 11 166) huoleen turvallisuudesta ja yksityisyydensuojasta tietoturvahyökkäysten yhteydessä. Monipuolisen datan, eli esimerkiksi potilaskertomusten ja geneettisten tietojen yhdistäminen voi puolestaan parantaa mallien ennustuskykyä (Garg ja muut, 2025). Procterin ja muiden (2023, s. 28) ajatus työkalupakista, johon kootaan menetelmiä luotettavan tekoälyn käyttäjävaatimusten määrittämiseksi, tukee konkreettisten työvälineiden käyttöä. Se yhdistäisi etnografiset havainnot suunnittelukysymyksiin, mukaan lukien luotettavan tekoälyn suunnittelumallien (design patterns) tunnistamisen. Esimerkiksi yksi suunnittelumalli voisi pyrkiä hahmottamaan moniammatillisten kokousten yhteisiä piirteitä, toisen mallin keskittyessä lääketieteellisten kuvien tulkintaan. Karimin ja muiden (2022, s. 54407) mukaan tilastollisesti merkittävät

ominaisuudet tunnistavaa white box -mallia voitaisiin käyttää tulkittavuuden tukena selittämässä miten löydökset vaikuttavat mallin tulokseen ja interaktioon.

Riskinä ja rajoitteena voidaan nähdä kirjallisuuskatsaus tutkimusmenetelmänä. Tutkimusaineistoksi valikoitui se mitä hakusanoilla tietystä tietokannasta löytyi, eikä mitään varmuutta ole siitä, ettei jokin oleellinen tutkimus olisi jäänyt tutkielmassa huomioimatta. Snyder (2019, s. 333) viittaa Tranfieldiin ja muihin, joiden mukaan perinteisen kirjallisuuskatsauksen riskinä on katsauksen pintapuolisuus ja epäsystemaattisuus. Tämä johtaa tiedonpuutteeseen siitä, mitä tutkimuskokoelma oikeastaan sanoo tai mihin se osoittaa. Riskinä on, että tekijät rakentavat tutkimuksen virheellisten olettamusten pohjalta. Tutkijoiden kerätessä valikoivasti tutkimuksen pohjalla olevaa aineistoa ja kenties jättäessään huomiotta epäyhtenäiset tutkimustulokset, voidaan kohdata vakavia ongelmia. Tutkimusta rajoittavana tekijänä voidaan nähdä myös tekoälyn valtava kehitysnopeus. Vaikka artikkelihaku rajattiin vuosille 2021–2024, erityisesti alkupään tutkimusaiheista on oletettavasti kertynyt sen jälkeen lisätietoa, tutkitut järjestelmät ovat kehittyneet ja laajentuneet, ja tässä työssä esiintyneet tiedot voivat jo työn julkaisuvaiheessa olla vanhentuneet.

Tutkimustulokset ovat soveltaen yleistettävissä myös muihin kriittisen infrastruktuurin tekoälyjärjestelmiin. El-Sappagh ja muut (2023, s. 11 274) viittaavat artikkelissaan González-Gonzaloon ja muihin sekä Li ja muihin, joiden mukaan luotettava tekoäly on merkittävä haaste turvallisuuskriittisillä aloilla, kuten lääketieteessä, oikeustieteessä, turvallisuusalalla. He viittaavat edelleen Kauriin ja muihin, joiden mukaan luotettavuus on välttämätön vaatimus myös muun muassa kaupallisella alalla, koulutuksessa, hallinnossa ja kodin automaatiossa.

Huomioitavaa on, että kehitys tekoälykentällä on valtavan nopeaa, joten uutta tutkimustietoa tarvitaan koko ajan. Esimerkiksi generatiivisen tekoälyn roolista, potentiaalista ja haasteista terveydenhuollossa tarvitaan lisätutkimusta. EU:n tekoälyasetus julkaistiin kesäkuussa 2024, joten sen vaikutuksista tekoälyn luotettavuuteen ei ole vielä ehtinyt

kertyä tutkimusnäyttöä. Tutkimusaineiston pohjalta korostui käsitys, että teoreettisen tiedon ja käytännön sovellusten välillä on toistaiseksi valtava kuilu. Empiiristä tutkimusta käytännön ympäristössä vaaditaan ehdottomasti lisää, jotta saadaan näyttöä luotettavan tekoälyn tilanteesta terveydenhuollon käyttökontekstissa.

Lähteet

Accenture. (2017). *Artificial intelligence: Healthcare's new nervous system. Executives for Health Innovation*. Noudettu 5.2.2025 osoitteesta https://www.ehidc.org/sites/default/files/resources/files/Accenture-Health-Artificial-Intelligence.pdf

Alsalem, M. A., Alamoodi, A. H., Albahri, O. S., Albahri, A. S., Martínez, L., Yera, R., Duhaim, A. M., & Sharaf, I. M. (2024). Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. *Expert Systems with Applications*, *246*, 123066. https://doi.org/10.1016/j.eswa.2023.123066

Bürger, V. K., Amann, J., Bui, C. K. T., Fehr, J., & Madai, V. I. (2024). The unmet promise of trustworthy AI in healthcare: Why we fail at clinical translation. *Frontiers in Digital Health*, *6*. https://doi.org/10.3389/fdgth.2024.1279629

Bærøe, K., Miyata-Sturm, A., & Henden, E. (2020). How to achieve trustworthy artificial intelligence for health. *Bulletin of the World Health Organization*, *98*(4), 257–262. https://doi.org/10.2471/BLT.19.237289

Dufva, M. (2024, tammikuuta 4). *Megatrendit 2024*. Sitra. https://www.sitra.fi/blogit/megatrendit-2024/

El-Sappagh, S., Alonso-Moral, J. M., Abuhmed, T., Ali, F., & Bugarín-Diz, A. (2023). Trustworthy artificial intelligence in Alzheimer's disease: State of the art, opportunities, and challenges. *The Artificial Intelligence Review*, *56*(10), 11149–11296. https://doi.org/10.1007/s10462-023-10415-5

Euroopan komissio. (2019). *Luotettavaa tekoälyä koskevat eettiset ohjeet*. Noudettu 15.1.2025 osoitteesta https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_FI.pdf

Euroopan parlamentin ja neuvoston asetus (EU) 2024/1689, Pub. L. No. 2024/1689 (2024). Noudettu 15.1.2025 osoitteesta https://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=OJ:L_202401689

Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, *6*, 01–11. https://doi.org/10.3389/fdgth.2024.1267290

Fehr, J., Jaramillo-Gutierrez, G., Oala, L., Gröschel, M. I., Bierwirth, M., Balachandran, P., Werneck-Leite, A., & Lippert, C. (2022). Piloting a Survey-Based Assessment of Transparency and Trustworthiness with Three Medical AI Tools. *Healthcare*, *10*(10), Article 10. https://doi.org/10.3390/healthcare10101923

Garg, P., Sharma, M. K., & Kumar, P. (2025). Transparency in Diagnosis: Unveiling the Power of Deep Learning and Explainable AI for Medical Image Interpretation. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-024-09896-5

Gundersen, T., & Bærøe, K. (2022). The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models. *Science and Engineering Ethics*, *28*(2), 17. https://doi.org/10.1007/s11948-022-00369-2

Herzog, C., Blank, S., & Stahl, B. C. (2024). Towards trustworthy medical AI ecosystems – a proposal for supporting responsible innovation practices in AI-based medical innovation. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-02082-z

Karim, M. R., Islam, T., Lange, C., Rebholz-Schuhmann, D., & Decker, S. (2022). Adversary-Aware Multimodal Neural Networks for Cancer Susceptibility Prediction From Multiomics Data | IEEE Journals & Magazine | IEEE Xplore. *IEEE Access*, *10*, 54386–54409. https://doi.org/10.1109/ACCESS.2022.3175816

Li, Y., Cai, X., Wu, C., Lin, X., & Cao, G. (2024). A Trustworthy Counterfactual Explanation Method With Latent Space Smoothing | IEEE Journals & Magazine | IEEE Xplore. *IEEE Transactions on Image Processing*, *33*, 4584–4599. https://doi.org/10.1109/TIP.2024.3442614

Liu, R., Rong, Y., & Peng, Z. (2020). A review of medical artificial intelligence. *Global Health Journal*, *4*(2), 42–45. https://doi.org/10.1016/j.glohj.2020.04.002

Mittal, S., Thakral, K., Singh, R., Vatsa, M., Glaser, T., Canton Ferrer, C., & Hassner, T. (2024). On responsible machine learning datasets emphasizing fairness, privacy

and regulatory norms with examples in biometrics and healthcare. *Nature Machine Intelligence*, *6*(8), 936–949. https://doi.org/10.1038/s42256-024-00874-y

Pourzolfaghar, Z., Alfano, M., & Helfert, M. (2023). Application of ethical AI requirements to an AI solution use-case in healthcare domain. *American Journal of Business*, *38*(3), 112–128. https://doi.org/10.1108/AJB-12-2022-0201

Procter, R., Tolmie, P., & Rouncefield, M. (2023). Holding AI to Account: Challenges for the Delivery of Trustworthy AI in Healthcare. *ACM Trans. Comput.-Hum. Interact.*, *30*(2), 31:1-31:34. https://doi.org/10.1145/3577009

Rathkopf, C., & Heinrichs, B. (2024). Learning to Live with Strange Error: Beyond Trustworthiness in Artificial Intelligence Ethics. *Cambridge Quarterly of Healthcare Ethics*, *33*(3), 333–345. https://doi.org/10.1017/S0963180122000688

Russell, & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson Education.

Sáez, C., Ferri, P., & García-Gómez, J. M. (2024). Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support. *Journal of Medical Internet Research*, *26*, e50295. https://doi.org/10.2196/50295

Sanmark, J., & Sanmark, E. (2024). *Mitä tiedämme generatiivisen tekoälyn hyödyistä terveydenhuollossa?* Noudettu 15.1.2025 osoitteesta https://www.duodecim-lehti.fi/xmedia/duo/duo18143.pdf

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The New System Technology*. Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Toscano, F., O'Donnell, E., Broderick, J. E., May, M., Tucker, P., Unruh, M. A., Messina, G., & Casalino, L. P. (2020). How Physicians Spend Their Work Time: An Ecological Momentary Assessment. *Journal of General Internal Medicine*, *35*(11), 3166–3172. https://doi.org/10.1007/s11606-020-06087-4

Tuomi, J. & Sarajärvi, A. (2018). Laadullinen tutkimus ja sisällönanalyysi. Tammi.

Vahdani, A. M., & Faghani, S. (2024). Deep Conformal Supervision: Leveraging Intermediate Features for Robust Uncertainty Quantification. *Journal of Imaging Informatics in Medicine*, 1–11. https://doi.org/10.1007/s10278-024-01286-5

Valtioneuvosto. (ei pvm.). Vahva ja välittävä Suomi: Pääministeri Petteri Orpon hallituksen ohjelma 20.6.2023. *Valtioneuvoston julkaisuja 2023:58*, *2023*. http://urn.fi/URN:ISBN:978-952-383-763-8

Zicari, R. V., Ahmed, S., Amann, J., Braun, S. A., Brodersen, J., Bruneault, F., Brusseau, J., Campano, E., Coffee, M., Dengel, A., Düdder, B., Gallucci, A., Gilbert, T. K., Gottfrois, P., Goffi, E., Haase, C. B., Hagendorff, T., Hickman, E., Hildt, E., … Wurth, R. (2021a). Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. *Frontiers in Human Dynamics*, *3*. https://doi.org/10.3389/fhumd.2021.688152

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., Holm, S., Kühne, U., Madai, V. I., Osika, W., Spezzatti, A., Schnebel, E., Tithi, J. J., Vetter, D., Westerlund, M., … Kararigas, G. (2021b). On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Frontiers in Human Dynamics*, *3*. https://doi.org/10.3389/fhumd.2021.673104

Liite 1. Selostus tekoälyn käytöstä tutkielmassa

Tekoälyä on käytetty tutkielmassa apuna kielenhuollossa sekä tekstin tiivistämisessä kaikissa pääluvuissa. Käytössä olleet kielimallit ovat olleet GPT-4o mini ja GPT-4.5.

Liite 2. Kirjallisuuskatsaukseen valitut artikkelit

ID	Tekijät	Vuosi	Artikkelin nimi	Julkaisu
1	Fehr, J. ja muut	2024	A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare	FRONTIERS IN DIGITAL HEALTH
2	Li, Y. ja muut	2024	A Trustworthy Counterfactual Explanation Method With Latent Space Smoothing	IEEE TRANSACTIONS ON IMAGE PROCESSING
3	Karim, MR. ja muut	2022	Adversary-Aware Multimodal Neural Networks for Cancer Susceptibility Prediction From Multiomics Data	IEEE ACCESS
4	Pourzolfaghar, Z. ja muut	2023	Application of ethical AI requirements to an AI solution use-case in healthcare domain	AMERICAN JOURNAL OF BUSINESS
5	Zicari, RV. ja muut	2021	Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier	FRONTIERS IN HUMAN DYNAMICS
6	Vahdani, AM. & Faghani, S	2024	Deep Conformal Supervision: Leveraging Intermediate Features for Robust Uncertainty Quantification	JOURNAL OF IMAGING INFORMATICS IN MEDICINE
7	Alsalem, MA. ja muut	2024	Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach	EXPERT SYSTEMS WITH APPLICATIONS
8	Procter, R. ja muut	2023	Holding AI to Account: Challenges for the Delivery of Trustworthy AI in Healthcare	ACM TRANSACTIONS ON COMPUTER-HUMAN INTERACTION
9	Zicari, RV. ja muut	2021	On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls	FRONTIERS IN HUMAN DYNAMICS
10	Mittal, S. ja muut	2024	On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare	NATURE MACHINE INTELLIGENCE
11	Fehr, J. ja muut	2022	Piloting a Survey-Based Assessment of Transparency and Trustworthiness with Three Medical AI Tools	HEALTHCARE
12	Gundersen, T. & Bæroe, K.	2022	The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models	SCIENCE AND ENGINEERING ETHICS
13	Bürger, VK. ja muut	2024	The unmet promise of trustworthy AI in healthcare: why we fail at clinical translation	FRONTIERS IN DIGITAL HEALTH
14	Herzog, C. ja muut	2024	Towards trustworthy medical AI ecosystems - a proposal for supporting responsible innovation practices in AI-based medical innovation	AI & SOCIETY
15	Garg, P. ja muut	2025	Transparency in Diagnosis: Unveiling the Power of Deep Learning and Explainable AI for Medical Image Interpretation	ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING