

Johan Sanmark ja Enni Sanmark

Systemoitu katsaus

Mitä tiedämme generatiivisen tekoälyn hyödyistä terveydenhuollossa?

Generatiivinen tekoäly on kuluneen vuoden aikana noussut yleiseen tietoisuuteen. Se on tekoälyn muoto, joka voi luoda täysin uutta sisältöä, kuten tekstiä, kuvaa, ääntä tai koodia sille annetun oppimaisaineiston perusteella. Generatiivisesta tekoälystä on kaavailtu ratkaisua terveydenhuollon globaaliin osaajapulaan. Tässä systemoidussa katsauksessamme selvitimme, minkälaista näyttöä generatiivisen tekoälyn hyödyistä terveydenhuollossa on. Vertailevia tutkimuksia generatiivisen tekoälyn käytöstä terveydenhuollossa on vain niukasti ($n = 9$). Tutkimusten perusteella generatiivinen tekoäly suoriutui vähintään käytössä olevan menetelmän tasoisesti dokumenttien luomisessa, kuvantamistutkimusten tulkinnaissa sekä ennustettaessa sairaalahoidon kestoa tai sairaalaan paluuta. Generatiivisen tekoälyn käyttöön liittyy kuitenkin riskejä, kuten tietosuojaa ja ammattilaisten liiallinen luottamus koneen tekemiin päätöksiin. Koska näyttö on niukkaa, tulevissa käyttöönotoissa on huolellisesti seurattava ja tutkittava käyttöönoton vaikutuksia.

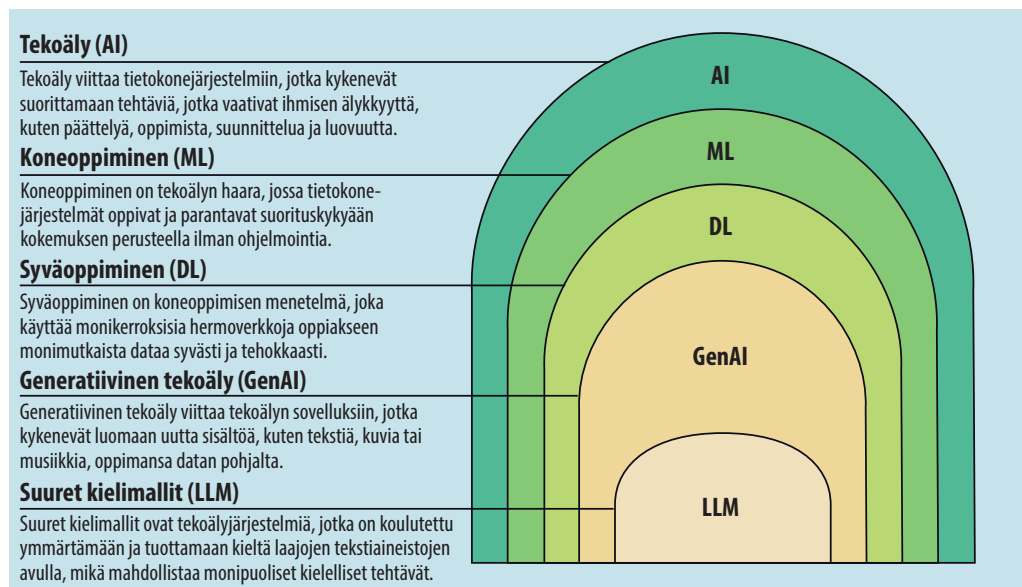
Generatiivinen tekoäly nousi laajalti ihmisten tietoisuuteen OpenAI:n julkaisemassa ChatGPT-palvelun marraskuussa 2022, vuosia aikaisemmin kuin oli odotettu. ChatGPT on vain yksi, vaikkakin edistynyt, esimerkki generatiivisesta tekoälystä, mutta se toi ensimmäistä kertaa teknologian jokaisen käyttöön. Yleisesti uskotaan, että generatiivinen tekoäly on tullut jäädäkseen ja muuttaa maailmaa samalla tavalla kuin tietokoneet, internet tai matkapuhelimet ovat tehneet, mikä lisää globaalisti bruttokansantuotetta (BKT) jopa 7 % (1).

Tekoäly voidaan määritellä tietokonejärjestelmänä, joka pystyy tehtäviin, jotka vaativat ihmisen älykkyyttä (2). Generatiivinen tekoäly on tekoälyn muoto, joka nimensä mukaan pystyy tuottamaan uutta materiaalia, kuten tekstiä, kuvaa, ääntä tai koodia (KUVA 1) (3). Generatiivinen tekoäly hyödyntää useita erilaisia teknologioita, joista yksi keskeinen ovat suuret kielimallit (LLM). Suuret kielimallit ovat syväoppimisen malleja, jotka ovat erikoistuneet luonnollisen eli ihmismäisen tekstin ymmärtä-

miseen ja tuottamiseen. Suuret kielimallit hyödyntävät vuonna 2017 kehitettyä transformer-teknologiaa oppiakseen valtavasta koulutusdatasta. Ne ennustavat tilastollisesti seuraavan sanan tai merkin ja muodostavat näin yhtenäistä tekstiä (4). Koulutustarpeen suuruusluokan hahmottamiseksi esimerkiksi GPT4-kielimalli on koulutettu datalla, joka vastaa noin 26 miljardia A4-sivua tekstiä (5).

ChatGPT:ssä käytössä olevan GPT-kielimallin lisäksi markkinoilla on jo runsaasti muitakin vastaavia kielimalleja, kuten Metan LLaMA tai Alfabetin Bard ja PALM (6). Näiden universaalien suurten kielimallien rinnalle on jo syntynyt myös kohdennetummalla ja suppeammalla aineistolla koulutettuja suuria kielimalleja, kuten NYU Langone Healthin NYUTron (7). Toinen esimerkki generatiivisen tekoälyn hyödyntämisestä teknologiasta ovat kuvantamistutkimuksissa hyödynnettävät GAN- ja diffuusiomallit (8).

Suomen julkinen sosiaali- ja terveydenhuolto on yhä syvenevässä kriisissä. Palvelutarve lisääntyy nopeasti, huoltosuhde heikkenee ja



AI = artificial intelligence, DL = deep learning, GenAI = generative artificial intelligence, LLM = large language models, ML = machine learning

KUVA 1. Tekoäly on kattotermi, joka viittaa tietojärjestelmiin, jotka kykenevät suorittamaan ihmisen älykkyyttä vaativia tehtäviä. Generatiivinen tekoäly ja sen osana suuret kielimallit erottuvat siinä, että ne voivat tuottaa täysin uutta, uniikkia sisältöä oppimansa perusteella.

yhteiskunnan maksukyvykyys tai -halukkuus ei pysy perässä. Tämä näkyy muun muassa voimakkaana henkilöstöpulana ja hoitoon pääsyn viivästymisenä yli lainsäädännön asettamien rajojen (9). Toisaalta jopa puolet sote-henkilöstön työajasta kuluu erilaiseen kirjaamiseen ja lausuntoihin (10). Näin ollen ei ole yllättävää, että taitavasti tekstiä käsittelevä generatiivinen tekoäly voisi mahdollisesti olla osa ratkaisua. Konsulttiyhtiö Accenture on arvioinut oman tutkimuksensa perusteella, että jopa 20 % sote-henkilöstön työstä on siirrettävissä generatiivisen tekoälyn hoidettavaksi (11). Myös vuonna 2023 tehdyssä hallitusohjelmassa on tunnistettu teknologian mahdollisuudet kirjaamalla ohjelmaan: ”Selvitetään todennäköisimmin automatisoitavissa olevat sosiaali- ja terveydenhuollon tehtävät sekä niihin liittyvät mahdollisuudet ja riskit” (12).

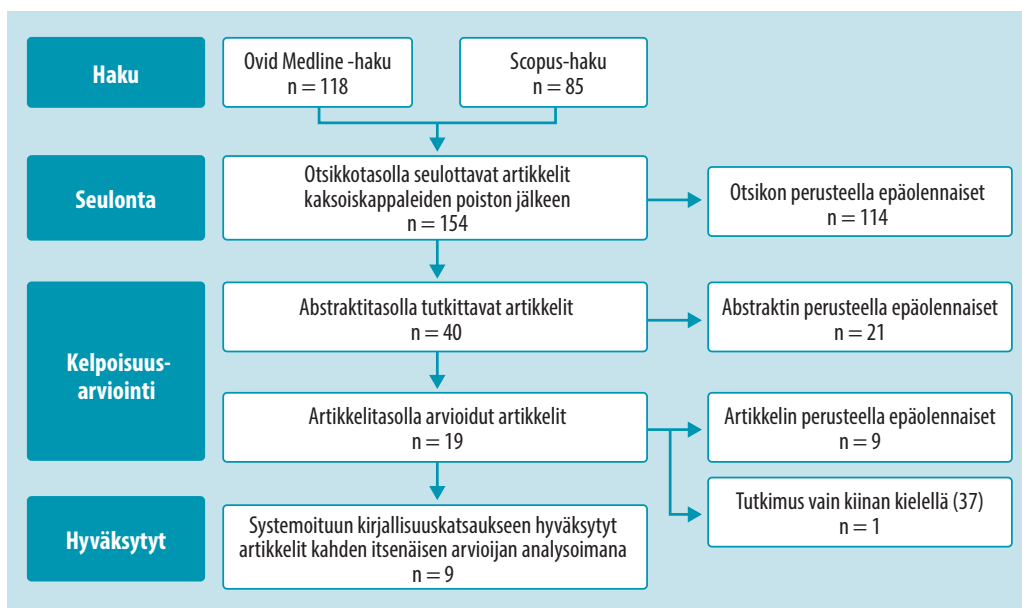
Generatiivisen tekoälyn käytöstä terveydenhuollossa on sekä kansallisesti että kansainvälisesti melko vähän kokemuksia ja vielä vähemmän tieteellisiä julkaisuja. Tieteellisen ja kaupallisen toiminnan myötä sen on kuitenkin esitetty pystyvän esimerkiksi seuraaviin tehtäviin: yhteenvedo ammattilaiselle keskeisistä po-

tilastiedoista, automaattiset sairauskertomusmerkinnät nauhoitetun vastaanoton perusteella, automaattinen lausuntojen laatiminen, räätilöityjen potilasohjeiden laatiminen, tiedon hakeminen tietokannoista, tiedolla johtaminen, riskipotilaiden tunnistaminen ja potilasviestinnän tehostaminen (13–19).

Generatiiviseen tekoälyyn kohdistuu valtavasti odotuksia kaikilla aloilla, myös terveydenhuollossa. Myös Suomessa generatiivisen tekoälyn käyttöönottoa sosiaali- ja terveydenhuollon organisaatioissa mietitään tai ainakin tulisi jo miettiä. Tämän systemoidun kirjallisuuskatsauksemme tarkoituksena oli selvittää, minkälaista näyttöä generatiivisen tekoälyn käytettävyydestä, hyödyistä ja mahdollisista riskeistä terveydenhuollon kontekstissa on jo olemassa.

Aineisto ja menetelmä

Haimme ja arvioimme tutkimuksia, joissa oli tutkittu generatiivisen tekoälyn käyttökohteita ja mahdollisia hyötyjä terveydenhuollossa. Katsauksessa noudatettiin PRISMA-ohjeistusta (20).



KUVA 2. Kaavio PRISMA-kriteerien mukaisesta systemaattisesta kirjallisuushausta ja artikkelivalinnasta.

Teimme 16.11.2023 tietokantahaun sekä Ovid Medline -tietokannasta sekä Scopus-tietokannasta samoilla hakutermeillä. Haku-termit olivat seuraavat: ("Generative AI" OR ChatGPT OR "Large language model") AND "health care" OR "Patient care") AND (result* OR outcome*). Haku rajattiin vertaisarvioituihin alkuperäisjulkaisuihin. Lisäksi aineistoon perehdyttäessä arvioitiin muita esiin tulleita julkaisuja, mutta näistä yksikään ei vastannut tutkimuskysymykseen, eikä niitä otettu mukaan tutkimukseen.

Lopulliseen analyysiin otettiin mukaan tutkimukset, jotka olivat alkuperäistutkimuksia, joissa käytettiin generatiivista tekoälyä, joissa tutkittiin todellisia terveydenhuollon tapauksia ja joissa generatiivisen tekoälyn kyvykkyyttä verrattiin tämänhetkisiin toimintamalleihin tai toisiin ohjelmistoihin. Lopulliseen analyysiin ei otettu mukaan pääkirjoituksia ja muita kirjoituksia, jotka eivät sisältäneet uusia tutkimustuloksia, muita ei-generatiivisia tekoälypohjaisia menetelmiä tutkivia tutkimuksia, tutkimuksia, joissa arvioitiin generatiivisen tekoälyn suoriutumista teoreettisissa tai keksityissä lääketieteellisissä kysymyksissä tai kokeissa taikka tutkimuksia, joissa pelkästään arvioitiin

generatiivisen tekoälyn vastauksia lääketieteellisiin kysymyksiin ilman vertailuryhmää.

Näitä valinta- ja poissulkukriteereitä käytettiin arvioitaessa hakuun sisältyneitä julkaisuja, ja niiden perusteella valikoitiin lopulliset tutkimukseen otetut julkaisut. Protokolla toteutettiin manuaalisesti niin, että tutkimukseen valittavia artikkeleita arvioi ja raportoi kaksi tutkijaa itsenäisesti. Erimielisyydet ratkaistiin keskustellen.

Tutkimukseen valituista tutkimuksista poimittiin ja analysoitiin asetelma, tutkimustyyppi, käytetty teknologia, käyttökohteet ja mittarit, generatiivisen tekoälyn suoriutuminen verrattuna verrokkiryhmään sekä kirjoittajien esiintuoma potentiaali ja riskit tulevaisuudessa. Tutkimuksia arvioitiin pääosin laadullisesti, eikä niistä tehty uusia tilastollisia analyyseja.

Tulokset

Lopulliseen analyysiin valittiin yhdeksän tutkimusta **KUVAN 2** protokollan mukaisesti. Jokaisesta tutkimuksesta analysoitiin tutkimusasetelma, tutkimustyyppi, käytetty teknologia, mahdollinen verrokkiryhmä, päätemuuttujan arviointiin käytetyt mittarit sekä tutkimustulos.

TAULUKKO. Tutkimukseen hyväksytyt artikkelit.

Tutkimus	Asetelma	Tutkimus- tyyppi	Teknologia	Mittari	Tulos
(28)	Arvioitiin LLM:n kykyä valita tilanteeseen sopiva kuvantamistutkimus (n = 27).	Takautuva vertaileva, ei-kontrolloitu tutkimus	GPT3.5 GlassAI	ACR AC (appropriateness criteria, tutkimuksen asianmukaisuuskriteeri)	Kielimallien välillä ei havaittu tilastollisesti merkitsevää eroa. Neuroradiologin suorituskkyky oli merkitsevästi parempi verrattuna sekä GPT:hen (p = 0,003) että GlassAI:hin (p = 0,013).
(22)	Arvioitiin LLM:n kykyä arvioida seuraavia asioita potilastapauksissa: 1) 30 vrk:n hoitoon paluun todennäköisyys (n = 20 takautuvasti ja n = 29 286 etenevästi) 2) Sairaalaakuolleisuus 3) Komorbiditeetti-indeksi 4) Sairaalaassaolon kesto 5) Vakuutuksen hylkäystodennäköisyys	Takautuva (kohdat 1–5) ja etenevä (kohta 1) kontrolloitu tutkimus	Oma kielimalli: NYUTron	AUC (area under the curve)	NYUTron-kielimallin AUC oli 78,7–94,9 % eli 5,4–14,7 % parempi tavanomaisiin ennustemalleihin verrattuna.
(24)	LLM käsitteli radiologin tekemiä lausuntoja thorax-TT:stä (62 kpl) ja aivojen magneettikuvauksesta (76 kpl). Arvioitiin kykyä 1) kääntää lausunnot maallikkokielelle, 2) antaa suosituksia lausuntojen perusteella ja 3) antaa toistuvasti samantasoiset käännökset.	Takautuva kokeellinen, kontrolloitu tutkimus	GPT3.5 GPT4.0	1) Viisipisteinen arvos- telujärjestelmä 2) Eri suositusten % 3) Yksi lausunto jaettiin 25 osaan ja arvioitiin kohta kohdalta astei- kolla ”hyvä-puuttuva- epäselvä-väärä”	1) Yleisarvosana 4,27/5 2) Yleisesti hyvät suositukset, 37 %:ssa tapauksista spesifiset löydöksiin liittyvät suositukset 3) Toistettavuus, GPT 3.5: 77 % hyviä ja GPT 4: 97 % hyviä
(29)	Arvioitiin LLM:n kykyä valita oikea diagnoosi ja hoito kynnär- ja olkapäävaurioissa (n = 29).	Etenevä kontrolloitu tutkimus	GPT3.0	Ortopedin arvio oi- keasta diagnoosista ja hoidosta	Kielimalli antoi oikean diagnoosin 93 %:ssa ja oikean hoidon 83 %:ssa tapauksista. Kuitenkin 46 %:ssa tapauksista hoitoehdotus oli osittain puutteellinen.
(21)	Vertailtiin LLM:iä, Adan, WebMD:n ja lääkärin tekemiä hoidon tarpeen arvioita (n = 37) ja top 3 -työdiagnooseja (n = 30).	Takautuva, kontrolloitu tutkimus	GPT3.5 GPT4.0	Vertailu lääkärin koti- utusvaiheen kirjauksiin (työdiagnosi) Vertailu lääkärin arvioon (hoidon tar- peen arvio)	Top 3 -työdiagnosien osuus ver- rattuna lääkärin kotiutusvaiheen kirjauksiin: Ada (63 %), GPT 3.5 (63 %), GPT 4.0 (50 %), WebMD, (57 %), lääkäri (69 %) Hoidon tarpeen arvion osuus ver- rattuna lääkärin arvioon: Ada (62 %) ChatGPT 3.5 (59 %) ChatGPT 4.0 (76 %), WebMD (70 %)
(26)	Vertailtiin LLM:n ja kirurgien kirjoittami- en suostumuslomakkeiden luettavuutta, tarkkuutta ja kattavuutta (n = 36).	Takautuva, kontrolloitu tutkimus	GPT3.5	Readability score	LLM loi luettavammat suostumuslo- makkeet, mutta ero ei ollut tilastolli- sesti merkitsevä (p = 0,10). Leikkaus- hyötyjen (p = < 0,001), leikkausvaih- toehtoien (p = < 0,001) ja yleiskuvan (p = 0,001) osalta LLM onnistui pa- remmin kuin kirurgi.
(23)	Arvioitiin LLM:n kykyä tulkita päivys- tyspoliklinikan keuhkoröntgenkuvia (n = 500).	Takautuva, ei-kontrolloitu tutkimus	Oma kie- limalli, ei nimeä	Viisipisteinen Likert- asteikko kokeneiden akuuttilääkäreiden arvioimana	Radiologin ja LLM:n välillä ei havaittu eroa tutkimusasetelmassa.
(27)	Arvioitiin LLM:n kykyä päätellä, tarvitsee- ko päivystyksessä tutkittu metastasoitu- nutta eturauhassyöpää sairastava potilas sairaalahoitoa (n=56).	Takautuva, kontrolloitu tutkimus.	GPT4.0	Vertailu kotiutusvai- heen kirjauksiin	LLM:n herkkyyks sairaalaan otettujen osalta 96 % ja tarkkuus kotiutettujen osalta 18 %. LLM teki saman työdiag- noosin lääkärin kanssa 88 %:ssa tapauksista mutta käytti tarkempaa terminologiaa (43 % vs 21 %, p = 0,02) ja kattavampaa diagnosilistausta (diagnoosien mediaanimäärä: 3 vs 2, p < 0,001).
(25) ¹	Arvioitiin tekoälyn (NLP = luonnollisen kielen käsittely) kykyä analysoida potilas- asiakirjamerkintöjä ja tunnistaa potilaat, joilla on jatkuvaa opioidien käyttöä suu- ren selkälölkauksen jälkeen (n = 965).	Takautuva, proof-of- concept- tyyppinen tutkimus	Abstraktissa mainittu vain NLP	Lääkärin arvio opioi- dien pitkäaikaisesta käytöstä (kyllä tai ei)	LLM tunnsti opioidien pitkäaikaiskäy- tön 96 %:ssa tapauksista ja käyttämät- ömyyden 86 %:ssa tapauksista.

¹ Pelkkä tiivistelmä käytettävissä

LLM = suuri kielimalli (large language model), TT = tietokonetomografia

Nämä esitetään **TAULUKOSSA**. Kaikkien analyysiin sisällytettyjen tutkimusten laatu arvioitiin riittäväksi.

Kielimallin kouluttaminen lääketieteellisellä aineistolla. Kaikissa systemoidun kat-
sausemme tutkimuksissa hyödynnettiin kielimallia. Kuudessa tutkimuksessa käytettiin GPT-kielimallia (versiot 3.5 tai 4). Näissä tutkimuksissa tutkijat huomauttivat, että kielimalli voisi suoriutua vielä paremmin lääketieteellisistä tehtävistä, jos se olisi koulutettu lääketieteellisellä datalla. Esimerkiksi hoidon tarpeen arviointiohjelmistoja vertailevassa tutkimuksessa lääketieteellisesti koulutetut ei-generatiiviset ohjelmistot suoriutuivat yleisesti hieman paremmin kuin ChatGPT, mutta nämä ohjelmistot oli koulutettu pelkästään hoidon tarpeen arviointitehtäviin (21). Oma, lääketieteelliseen toimintaan käytetty kielimallia käytettiin kahdessa tutkimuksessa, ja näissä kielimallin havaittiin suoriutuvan ihmisen veroisesti tai aiempia menetelmiä paremmin (22,23).

Kielimallit apuna dokumentaatioissa ja tiedonhaussa. Dokumentaatioon keskittyvissä tutkimusasetelmissä havaittiin, että promptin (kielimallille annettava komento) optimointi paransi kielimallin tarkkuutta tehtävissä (24). Lisäksi kielimallien havaittiin toimivan hyvin juuri ei-rakenteisen tekstin arvioimisessa (25). Koska kielimallit pystyvät generoimaan esimerkiksi suostumusdokumenteja jopa laadukkaammin kuin lääkärit, kirjoittajat päättelivät, että tällaiset tehtävät voitaisiin jatkossa siirtää lääkäreiltä tekoälylle (26).

Mahdolliset riskit. Analyysiin mukaan ote-
tuissa tutkimuksissa ei ilmennyt generatiivisen tekoälyn käyttöön liittyviä, käyttöä ehdottomasti estäviä merkittäviä riskejä. Esimerkiksi hoidon tarpeen arvioinnissa liian uskaliaita ("unsafe") päätöksiä GPT4 ei tehnyt lainkaan enempää verrattuna hoidon tarpeen arviointiin käytettyihin apuohjelmistoihin. Ammattilaisten tekemiä päätöksiä pidettiin tutkimuksessa standardina (21). Generatiivisen tekoälyn on havaittu hallusinoivan eli keksivän tietoa, jota ei ole olemassa, joitakin epätyypillisimmin käytettyjä lyhenteitä käytettäessä (27). Lisäksi on huomautettu, että vaikka GPT3 antoi oikeat hoito-ohjeet, sen ohjeet olivat lähes puo-

lessa tapauksista puutteelliset. Se ei myöskään osannut luotettavasti päättää leikkaushoidon ja konservatiivisen hoidon väliltä. Lisäksi tutkimusryhmä huomautti, että kielimallin kohtamiseen ei sisälly samaa empatiaa tai vuorovaikutusta kuin ihmisten välillä, eikä kielimalli voi näin korvata ihmistä (21).

Tehtäviä useilla erikoisaloilla. Systemoituun katsaukseen sisällytetyistä tutkimuksista kolme oli tehty päivystyspoliklinikassa (21,23,27), kaksi radiologian osastolla (24,28), kolme kirurgian osastolla (25,26,29) ja yksi sairaalan organisaatioissa (28). Muilta aloilta kliinisiä, vertailevia tutkimuksia generatiivisen tekoälyn hyödyntämisestä ei vielä ole raportoitu.

Pohdinta

Kirjallisuuskatsauksessamme havaitsimme, että käytännön tutkimuksia ja näyttöä generatiivisen tekoälyn hyödyistä on vielä niukasti ja vain valikoiduissa tapauksissa. Osa tutkimuksista oli tehty jo vanhentuneilla kielimalleilla. Generatiivisella tekoälyllä on kuitenkin tutkimusten perusteella merkittävää potentiaalia esimerkiksi ennustettaessa sairaalahoidon kestoa, kuolleisuutta tai sairaalaan paluuta, dokumenttien kuten suostumusten luomisessa sekä radiologisten kuvantamistutkimusten tulkin-
nassa (22,23,26). Näissä kaikissa tapauksissa generatiivinen tekoäly suoriutui vähintään yhtä hyvin tai jopa paremmin kuin nykyisin käytössä oleva menetelmä. Muissa sovelluksissa, kuten hoidon tarpeen arvioinnissa, radiologisten kuvantamistutkimusten valinnassa ja maallikkokielelle kääntämisessä sekä yläraajaortopedisen diagnoosin ja hoitolinjan valinnassa kielimalli todettiin käyttökelpoiseksi, mutta se suoriutui tehtävistä huomoinnakin kokenut ammattilainen (21,24,28).

Generatiivisen tekoälyn käyttöön liittyviä riskejä on pohdittu laajasti kirjallisuudessa. Riskeiksi on nähty esimerkiksi hallusinointi, puutteellinen tietosuojajärjestelmä ja ammattilaisten liiallinen nojautuminen koneen tekemiin päätöksiin (19,30,31). Hallusinaatiot eli kielimallien tuottamat epätodet vastaukset nähdään yhtenä merkittävimmistä riskeistä terveydenhuollon

Ydinasiat

- ▶ Generatiivinen tekoäly muuttaa myös terveydenhuollon toimintatapoja.
- ▶ Tutkimusnäyttöä generatiivisen tekoälyn hyödyistä terveydenhuollossa on toistaiseksi vain vähän.
- ▶ Julkaistujen tutkimusten perusteella generatiivinen tekoäly terveydenhuollossa tuo lisäarvoa esimerkiksi tekstien laatimiseen, radiologiseen tulkintaan ja ennusteiden tekemiseen.
- ▶ Generatiivinen tekoäly on todennäköisesti osa terveydenhuollon kriisin ratkaisua, mutta rajallisen näytön vuoksi käyttöön-otot on tehtävä harkiten ja hyötyjä on tutkittava lisää.

ympäristössä (19). Jaettua tietoa siitä, miten hallusinaatioista voitaisiin päästä eroon, ei vielä ole.

Tietosuoja ja luottamusriskit liittyvät merkittävästi siihen, että markkinoita hallitsevat nykyisin kaupallisten toimijoiden kielimallit, joiden lähdekoodiin tai koulutusaineistoihin ei ole näkyvyyttä. Läpinäkyvyysongelma johtuu siitä, että tekoäly käyttää tyypillisesti syväoppimisen (deep learning) algoritmeja sisällön tuottamiseen. Syväoppimisen mallit ovat niin sanottuja mustia laatikoita, joissa käytetään monia kerroksia monimutkaista algoritmista päättelyä, mikä tekee niiden vastausten syntyvän ymmärtämisen käyttäjille ja jopa kehittäjille mahdottomaksi tai erittäin vaativaksi (32).

Kaupallisissa kontekstissa on havaittu, että ihmiset luottavat sokeammin sellaisiin koneen tekemiin päätöksiin, joiden päätöksentekologiikkaa he eivät tunne (33). Tietosuoja- ja luottamusriskeihin ei otettu kantaa systemoituun katsaukseen valikoiduissa tutkimuksissa, vaan niissä pyrittiin selvittämään kielimallien kyvykkyyttä tehtävissä, joista ovat yleensä vastanneet ihmiset. Jatkossa tutkimuksia tulisi myös lääketieteessä tehdä siitä, miten generatiivisen tekoälyn käyttö vaikuttaa terveydenhuollon ammattilaisten toimintaan ja päätöksentekoon.

Jokaisen terveydenhuollon käyttökohteen osalta tulisi tehdä kattava riskinarviointi.

Tutkimuksissa käytettiin pääasiassa joko kaupallista, universaalista GPT-kielimallia tai vaihtoehtoisesti terveydenhuollon käyttöön erityisesti koulutettuja kielimalleja, kuten NYUTronia (22). Erityisesti lääketieteellisellä aineistolla koulutetut kielimallit vaikuttivat suoriutuvan paremmin kuin ei-lääketieteellisellä aineistolla koulutetut universaalit kielimallit. Otanta on kuitenkin pieni, eikä tämän perusteella voida tehdä yleistyksiä. Useammassa GPT-kielimallia käyttäneessä tutkimuksessa myös pohdittiin, voisiko kielimallin kyvykkyys vielä parantua, jos se koulutettaisiin lääketieteellisellä aineistolla (21,24,28). Tämä on kysymys, johon myös suomalaisessa terveydenhuollossa pitää pystyä lähivuosina vastaamaan: missä tehtävissä voidaan käyttää tehokkaita, universaaleja kielimalleja, missä niiden käyttömahdollisuuksia voidaan laajentaa RAG-tekniikalla (retrieval augmented generation) ja mihin tehtäviin vaaditaan tarkasti valitulla, tunnetulla aineistolla koulutettu kielimalli (34).

Tekoälylle ja sitä myöten generatiiviselle tekoälylle ei toistaiseksi ole olemassa omaa lainsäädäntöään. Euroopan komissiossa on kuitenkin valmisteilla tekoälyasetus (AI Act), mutta sen voimaantulon ajankohta ei vielä ole tiedossa (35). Asetuksen valmistelutyö on aloitettu jo ennen generatiivisen tekoälyn ilmiömaista kehittymistä kuluneen vuoden aikana, joten sitä, miten lainsäädäntö ottaa jatkossa huomioon generatiivisen tekoälyn käyttämisen esimerkiksi lääketieteessä, ei vielä varmuudella tiedetä. Nykyisten luonnosten perusteella valtaosa käyttötapauksista terveydenhuollon viitekehityksessä luokiteltaneen suuren riskin käyttökohteiksi (36).

Generatiivisen tekoälyn käyttöönotto lääketieteellisissä tehtävissä edellyttäne Suomessa pääosin monimutkaisen ja osin päällekkäisen sääntelyn huomioimista: tekoälyasetus (AI Act), tietosuoja-asetus (GDPR), lääkinnällisten laitteiden asetus (MDR), asetus eurooppalaisesta terveystiedon avaruudesta (EHDS) ja toisiolaki. Sääntelyä ei mainittu tämän katsauksemme tutkimuksissa, mikä liittyy siihen, että yksikään tutkimus ei ollut EU-alueelta.

Systemoidun katsauksemme perusteella kliinisiä käyttökohteita generatiivisen tekoälyn hyödyntämiselle on toistaiseksi tutkittu vain yksittäisillä erikoisaloilla ja yksittäisissä käyttökohteissa. Yhdessäkin tutkimuksessa ei esimerkiksi tutkittu generatiivista tekoälyä potilasasiakirjamerkintöjen kirjoittamisessa, tiivistämisessä tai lausuntojen automatisoinnissa, vaikka nämä käyttökohteet on nähty ilmeisinä juuri terveydenhuollossa, jossa henkilöstöpula on globaali ongelma (11). Generatiivinen tekoäly tulisi tuoda osaksi suomalaista sosiaali- ja terveydenhuoltoa jo ennen kattavan tieteellisen näytön kertymistä. Käyttökohteet on valittava huolellisesti teknologian rajoitukset huomioiden, riskinarviointien tulee olla riittävän kattavia ja tulokset tulisi dokumentoida ja julkaista.

Vaikka pelkästään PubMedistä löytyy tuhansia generatiivista tekoälyä ja ChatGPT:tä käsitteleviä julkaisuja, vain harva tutkimus on tehty oikeassa terveydenhuollon ympäristössä ja realistisessa asetelmassa sekä sisältänyt vertailuryhmän. Siksi myös tähän systemoituun katsaukseemme valikoitui vain yhdeksän julkaisua. Kaikki katsauksen julkaisut ovat lyhyeltä ajanjaksolta touko-lokakuulta 2023,

joten on odotettavissa, että tieto lisääntyy nopeasti. Tämä katsaus tehtiin tietoisesti jo nyt, sillä generatiivisen tekoälyn käyttöönotto tai käyttöönoton harkinta on käynnissä suomalaisessa terveydenhuoltojärjestelmässä. Tämän harkinnan tukena tulisi hyödyntää olemassa olevaa tutkittua tietoa.

Lopuksi

Generatiivinen tekoäly mullistaa jo nykyisen tutkimustiedon valossa terveydenhuollon toimintatavat ja sen myötä koko terveydenhuoltojärjestelmän. Ensimmäiset tutkimustulokset terveydenhuollon kontekstissa ovat lupaavia, mutta laadukkaita, vertailevia tutkimuksia generatiivisen tekoälyn käytöstä terveydenhuollossa on vielä vähän. Generatiivinen tekoäly otettaneen osaksi suomalaista terveydenhuoltojärjestelmää jo ennen kuin merkittävää tutkittua näyttöä sen hyödyistä on käytettävissä. Siksi pidämme tärkeänä, että uusissa generatiivisen tekoälyn käyttöönotoissa tutkittaisiin ja raportoitaisiin hyötyjä vähintään kansallisella tasolla. ■

JOHAN SANMARK, LL, terveydenhuollon erikoislääkäri, TKI-johtaja

Länsi-Uudenmaan hyvinvointialue

ENNI SANMARK, LL, terveydenhuollon erikoislääkäri, terveydenhuollon dosentti, tutkimusryhmän johtaja
HUS ja Helsingin yliopisto

VASTUUTOIMITTAJA

Tuomas Mirtti

SIDONNAISUUDET

Johan Sanmark: Luentopalkkio/asiantuntijapalkkio (Astra Zeneca, Boehringer Ingelheim, Validia, Zedocan), luottamustoimet (Nuorten Lääkärien Yhdistys, Lääkäriliitto)

Enni Sanmark: Luentopalkkio/asiantuntijapalkkio (Boehringer Ingelheim, Pfizer, Roche, B. Braun, Abbvie, Orion, Terveystalo, Lääkäriliitto), korvaukset koulutus- ja kongressikuluista (Roche), luottamustoimet (Nuorten Lääkärien Yhdistys, Duodecim)

KIRJALLISUUTTA

- Generative AI could raise global GDP by 7%. Goldman Sachs Research, julkaistu verkossa 5.4.2023. www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html.
- Sheikh H, Prins C, Schrijvers E. Artificial intelligence: definition and background. Kirjassa: Mission AI. Research for Policy. Dordrecht: Springer 2023, s.15–6.
- Feuerriegel S, Hartmann J, Janiesch C, ym. Generative AI. Bus Inf Syst Eng 2024;66:111–26.
- Vaswani A, Shazeer N, Parmar N, ym. Attention is all you need. ArXiv, julkaistu verkossa 12.6.2017. DOI:10.48550/arXiv.1706.03762.
- Research. GPT-4. OpenAI 2023. <https://openai.com/research/gpt-4>.
- Naveed H, Khan A, Qiu S, ym. A comprehensive overview of large language models. ArXiv, julkaistu verkossa 2.11.2023. DOI:10.48550/arXiv.2307.06435.
- Jiang LY, Liu XC, Nejatian NP, ym. Health system-scale language models are all-purpose prediction engines. Nature 2023;619:357–62.
- Khader F, Müller-Franzes G, Tayebi Arasteh S, ym. Denoising diffusion probabilistic models for 3D medical image generation. Sci Rep 2023;13:7303.
- Sosiaali- ja terveysalan henkilöstön riittävyiden ja saatavuuden turvaaminen. Tiekartta 2022–2027. Sosiaali- ja terveysministeriön julkaisuja 2023:8. Helsinki: STM 2023. https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/164634/STM_2023_8.pdf.
- Toscano F, O'Donnell E, Broderick JE, ym. How physicians spend their work time: an ecological momentary assessment. J Gen Intern Med 2020;35:3166–72.
- Artificial intelligence: healthcare's new nervous system. Dublin: Accenture 2017. www.accenture.com/content/dam/accnture/final/a-com-migration/manual/r3/pdf/pdf-49/Accenture-health-artificial-intelligence-j.pdf.
- Vahva ja välittävä Suomi – hallituksen visio. Pääministeri Petteri Orpon hallituksen ohjelma. Valtioneuvoston julkaisuja 2023:58. <https://valtioneuvosto.fi/hallitukset/hallitusohjelma#/>.
- Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health, julkaistu verkossa 6.2.2023. DOI:10.1016/S2589-7500(23)00021-3.
- Choi HS, Song JY, Shin KH, ym. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023;3:209–16.
- AWS HealthScribe features. Amazon AWS 2023. <https://aws.amazon.com/healthscribe/features/>.
- Aljindan FK, Shawosh MH, Altamimi L, ym. Utilization of ChatGPT-4 in plastic and reconstructive surgery: a narrative review. Plast Reconstr Surg Glob Open, julkaistu verkossa 26.10.2023. DOI:10.1097/GOX.0000000000005305.
- Amin K, Khosla P, Doshi R, ym. Artificial intelligence to improve patient understanding of radiology reports. Yale J Biol Med 2023;96:407–17.
- Benirschke RC, Wodskow J, Prasai K, ym. Assessment of a large language model's utility in helping pathology professionals answer general knowledge pathology questions. Am J Clin Pathol 2024;161:42–8.
- Clusmann J, Kolbinger FR, Muti HS, ym. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3:141.
- Page MJ, McKenzie JE, Bossuyt PM, ym. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, julkaistu verkossa 29.3.2021. DOI:10.1136/bmj.n71.
- Fraser H, Crossland D, Bacher I, ym. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR Mhealth Uhealth, julkaistu verkossa 3.10.2023. DOI:10.2196/49995.
- Jiang LY, Liu XC, Nejatian NP, ym. Health system-scale language models are all-purpose prediction engines. Nature 2023;619:357–62.
- Huang J, Neill L, Wittbrodt M, ym. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Netw Open, julkaistu verkossa 2.10.2023. DOI:10.1001/jamanetworkopen.2023.36100.
- Lyu Q, Tan J, Zapadka ME, ym. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6:9.
- Seng EC, Mehdipour S, Simpson S, ym. Tracking persistent postoperative opioid use: a proof-of-concept study demonstrating a use case for natural language processing. Reg Anesth Pain Med 2023. DOI:10.1136/rapm-2023-104629.
- Decker H, Trang K, Ramirez J, ym. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open, julkaistu verkossa 2.10.2023. DOI:10.1001/jamanetworkopen.2023.36997.
- Gebräel G, Sahu KK, Chigara B, ym. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. Cancers (Basel) 2023;15:3717.
- Nazario-Johnson L, Zaki HA, Tung GA. Use of large language models to predict neuroimaging. J Am Coll Radiol 2023;20:1004–9.
- Daher M, Koa J, Boufadel P, ym. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int 2023;7:2534–41.
- Jeyaraman M, Balaji S, Jeyaraman N, ym. Unraveling the ethical enigma: artificial intelligence in healthcare. Cureus, julkaistu verkossa 10.8.2023. DOI:10.7759/cureus.4326.
- Duffourc M, Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients. JAMA 2023;330:313–4.
- Babic B, Gerke S, Evgeniou T, ym. Beware explanations from AI in health care. Science 2021;373:284–6.
- Stefano T, Kellogg K, Menietti M, ym. Why providing humans with interpretable algorithms may, counterintuitively, lead to lower decision-making performance. MIT Sloan Research Paper 2022;6797.
- Lewis P, Perez E, Piktus A, ym. Retrieval-augmented generation for knowledge-intensive NLP tasks. ArXiv 2020;2005.11401.
- Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts (AI Act). Bryssel: European Commission 2021. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206>.
- Duffourc MN, Gerke S. The proposed EU directives for AI liability leave worrying gaps likely to impact medical AI. NPJ Digit Med 2023;6:77.
- Haoran X, Xiaoyan C, Huiming Z, ym. Construction and practice of big data platform for self-monitoring and follow-up of patients after artificial mechanical valve replacement with chatGPT. Chinese J Pract Nursing 2023;39:2276–84.