WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Tekoälyagentti hoiti kokonaisia potilastapauksia itsenäisesti — simulaattorissa, aiempien potilastietojen perusteella

MIRA on uudenlainen lääketieteellinen tekoäly: yhden kysymyksen vastaamisen sijaan se käy läpi kokonaisen potilastapauksen simuloidussa sairaalan potilastietojärjestelmässä — ottaa anamneesin, määrää ja tulkitsee tutkimuksia, päätyy diagnoosiin ja laatii määräykset. Julkisesta tietokannasta poimituissa 574 retrospektiivisessä tapauksessa, jotka kattoivat kahdeksan ennalta valittua diagnoosia, kirjoittajat raportoivat sen päihittäneen lääkärit diagnostisessa tarkkuudessa ja tehneen suurelta osin hoitosuositusten mukaisia ja lääkitysturvallisia päätöksiä. Jokainen tämän virkkeen rajaus on tärkeä: järjestelmä toimi tekstipohjaisessa hiekkalaatikossa menneillä potilastiedoilla; suuri osa sen etumatkasta tuli sairauksista, joissa tutkimustulokset olivat yksiselitteisiä; se hyödynsi noin kaksi kertaa enemmän verikokeita kuin lääkärit; ja useita tuloksia pisteytettiin sen mukaan, mitä alkuperäiseen potilaskertomukseen oli kirjattu. Todellinen edistysaskel on agentti, joka toimii koko työnkulun läpi — ei todiste siitä, että kone diagnosoisi nyt lääkäriä paremmin. Kirjoittajat sanovat tämän itse: yleistettävyyttä, turvallisuutta ja hallintaa on vielä tutkittava prospektiivisesti todellisessa kliinisessä ympäristössä.

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Towards autonomous medical artificial intelligence agents*, Dyke Ferber, Lars Hilgers, Christiane Höper and colleagues; senior author Jakob Nikolas Kather, Nature (2026) · DOI: 10.1038/s41586-026-10675-5

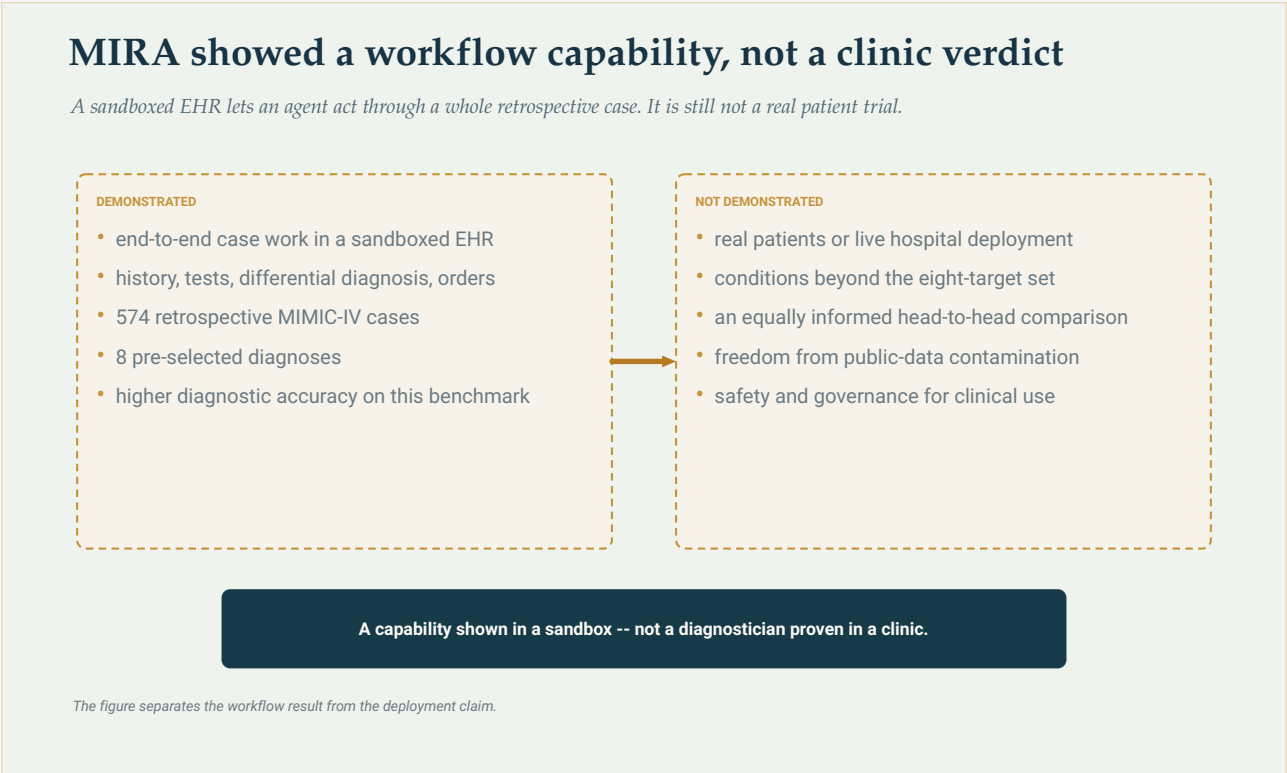
Kysymykseen vastaaminen ei ole sama asia kuin koko potilastapauksen hoitaminen

Useimmat lääketieteellistä tekoälyä koskevat jutut käsittelevät mallia, joka *vastaa*. Sille annetaan potilastapaus tai tenttikysymys, ja se palauttaa diagnoosin tai kappaleen neuvoja ja saa hyvän pistemäärän. Hyödyllistä, mutta rajallista: kuin fiksu konsultti, joka ei koskaan koske potilaskertomukseen.

MIRA on rakennettu tekemään jotain muuta, ja juuri tämä ero on varsinainen uutinen. Yhteen kysymykseen vastaamisen sijaan se käy läpi koko tapauksen: lukee potilaan tiedot, päättää mitä anamneesista vielä tarvitaan, määrää laboratorio-, kuvantamis- ja mikrobiologisia tutkimuksia, lukee tulokset, rajaa erotusdiagnoosia ja kirjoittaa sen jälkeen tarvittavat määräykset — reseptit, sairaalaanoton, lähetteen leikkaukseen. Se tekee tämän toimimalla *potilastietojärjestelmän sisällä* kliinikon tavoin sen sijaan, että se tuottaisi vapaata tekstiä ihmisen siirrettäväksi järjestelmään.

Se on todellinen askel: neuvovasta järjestelmästä järjestelmään, joka toimii koko työnkulun läpi. Ja juuri tällainen askel litistyy uutisoinnissa helposti muotoon ”tekoäly päihittää lääkärit”. Siksi kannattaa olla täsmällinen siitä, mitä testattiin ja missä.

Yhden virkkeen versio: MIRA hoiti kokonaisia tapauksia itsenäisesti ja päihitti tässä vertailussa lääkärit diagnostisessa tarkkuudessa — mutta se teki sen hiekkalaatikossa, retrospektiivisillä potilastiedoilla, kahdeksan ennalta valitun diagnoosin joukossa, vain tekstin välityksellä, ja suuri osa sen etumatkasta tuli sairauksista, joissa tutkimustulokset olivat kaikkein yksiselitteisimpiä. Edistysaskel on todellinen. Tulostaulu ei ole klinikka.



MIRA osoitti pystyvänsä käymään koko työnkulun läpi hiekkalaatikkoon rajatussa potilastietojärjestelmässä. Se ei osoittanut todellista kliinistä käyttöönottoa, laajaa diagnostista kattavuutta eikä reilua vertailua lääkäreihin tilanteessa, jossa molemmilla osapuolilla olisi yhtä paljon tietoa.

Original Aurora diagram — The Clean Paper CC BY 4.0

Mitä kirjoittajat tekivät

MIRA — Medical Intelligence for Reasoning and Action — on OpenAI:n malleihin perustuva autonominen agentti: keskusteleva ja toimiva osa käyttää **GPT-4o**-mallia, ja erillinen suunnitteluvaihe käyttää OpenAI:n **o1**-päättelymallia. Ne toimivat räätälöidyssä, standardien mukaisessa ympäristössä, joka perustuu HL7 FHIR -standardiin — samaan tietostandardiin, jolla todelliset sairaalat siirtävät potilastietoja — ja jossa on yli kahdeksankymmentätuhatta mahdollista kliinistä toimintoa. Hiekkalaatikossa MIRA voi hakea potilaan historian, määrätä ja tulkita laboratoriokokeita, kuvantamista ja mikrobiologiaa, muodostaa erotusdiagnoosin ja laatia hoitosuunnitelmia: määrätä lääkkeitä, ajoittaa kirurgisia toimenpiteitä ja suunnitella sairaalaanottoa. Yksi yksityiskohta on syytä nostaa esiin heti: MIRA:n vastauksia pisteyttäneet automaattiset ”tuomarit” olivat itse GPT-4o-malleja — samaa malliperhettä kuin testattava järjestelmä. Kun vertaisarvioija nosti tämän ongelman esiin, kirjoittajat täydensivät arviointia erikoislääkärin tarkistuksella.

Testiaineisto tuli **MIMIC-IV**-tietokannasta, joka on suuri, julkisesti saatavilla oleva anonymisoitu sähköisten potilastietojen aineisto. Noin 300,000 Beth Israel Deaconess Medical Centerissä vuosina 2008–2019 hoidetusta potilaasta kirjoittajat valitsivat **574 potilastapausta**, jotka kattoivat **kahdeksan kohde-diagnoosia** — vatsan alueen sairauksia ja sisätautien päivystyksiä, muun muassa umpilisäketulehduksen, haimatulehduksen, keuhkokuumeen ja virtsatieinfektion. Jokaisessa tapauksessa MIRA aloitti rajallisista tiedoista ja joutui päättämään askel askeleelta, mitä tehdä seuraavaksi — samanmuotoinen prosessi kuin todellinen tutkimusketju, mutta aineistona oli potilaskertomus, jonka todellinen lopputulos oli jo tiedossa.

Sen suoriutumista verrattiin samoissa tapauksissa lääkäreihin.

Mitä ”autonominen agentti hiekkalaatikkoon rajatussa potilastietojärjestelmässä” oikeastaan tarkoittaa

Tässä kolme sanaa kantaa paljon merkitystä, joten ne kannattaa avata.
Agentti tarkoittaa, ettei järjestelmä ole pelkkä kehotteeseen vastaava chatbot. Se toimii silmukassa: tarkastelee nykytilannetta, valitsee toiminnon (määrää tämän tutkimuksen, kysy tuosta anamneesitiedosta), näkee tuloksen ja valitsee seuraavan toiminnon — kunnes se päätyy diagnoosiin ja suunnitelmaan. ”Älykkyys” on kielimalli; *toimijuus* syntyy sen ympärille rakennetusta järjestelmästä, joka muuntaa mallin tekstin sallituiksi potilastietojärjestelmän toiminnoiksi ja syöttää tulokset takaisin mallille.
Hiekkalaatikkoon rajattu potilastietojärjestelmä tarkoittaa simuloitua, eristettyä kopiota potilastietojärjestelmästä, ei toimivan sairaalan järjestelmää. MIRA voi ”määrätä” tutkimuksen ja saada sen tuloksen, jonka kyseinen potilas todella sai, koska tapaus on historiallinen ja

vastaus on jo kirjattu. Mikään sen toiminta ei koske oikeaa potilasta tai todellista klinikkaa. **Autonominen** tarkoittaa, että se vie tapauksen alusta loppuun ilman ihmistä päätöksentekosilmukassa — *hiekkalaatikon sisällä*. Se **ei** tarkoita, että järjestelmä olisi päästetty hoitamaan potilaita ilman valvontaa. Nämä ovat hyvin eri väitteitä, ja vain ensimmäistä testattiin.
Hiekkalaatikosta vielä yksi asia, koska se on helppo kuvitella väärin: MIRA voi *määrätä* minkä tahansa haluamansa tutkimuksen, mutta järjestelmä voi palauttaa tuloksen vain, jos kyseinen tutkimus **todella tehtiin** tälle potilaalle. Jos se pyytää veripaneelia, joka potilaalta oikeasti otettiin, se saa todelliset arvot; jos se pyytää kuvantamista, jota kukaan ei määrännyt, vastaukseksi tulee ”N/A — could not be performed”, ei keksittyä lukua. Anamneesi toimii samalla tavalla: jos potilaskertomuksessa ei ole tietoa, jota MIRA kysyy, simuloitu potilas sanoo yksinkertaisesti, ettei tiedä. MIRA ei siis voi taikoa esiin yhtä ratkaisevaa tutkimusta, jota ei koskaan tehty — se voi käyttää vain sitä, mitä todellinen tutkimusketju sattui sisältämään.
Erottelu on tärkeä, koska otsikoiden sana ”autonominen” tulkitaan kaikkein helpoimmin jälkimmäisellä tavalla.

Mitä he löysivät

Vertailussa **MIRA päihitti lääkärit diagnostisessa tarkkuudessa** — mutta otsikko peittää alleen kolme eri lukua, jotka on helppo sekoittaa, joten ne kannattaa erottaa toisistaan.

Kaikissa **574 tapauksessa** MIRA nimesi oikean diagnoosin **88.9%**:ssa tapauksista, kun vertailukohtana oli kullekin MIMIC-IV-potilaalle kirjattu kotiutusdiagnoosi. Suorassa vertailussa ihmisiin lääkärit eivät käsitelleet kaikkia 574 tapausta, vaan yhteisen **311 tapauksen osajoukon**, käyttäen samoja potilastietoja ja työkaluja kuin MIRA. Näissä samoissa 311 tapauksessa MIRA sai tulokseksi **87.8%**, ja sitä verrattiin kahteen erikseen rekrytoituun ryhmään. Ensimmäinen oli **kokeneiden ryhmä**: neljä erikoislääkäriä, joilla oli 7–11 vuoden kokemus ja joiden tulos oli **78.1%**. Toinen oli **eri kokemustasojen ryhmä**: enimmäkseen nuoria erikoistuvia lääkäreitä sekä kaksi erikoislääkäriä, mikä muistuttaa paremmin todellisen päivystyksen henkilöstörakennetta; heidän tuloksensa oli **71.1%**. Samat tapaukset, samat työkalut — ja järjestys oli tekoäly ensimmäisenä, kokeneet lääkärit toisena ja nuorempiin painottunut ryhmä kolmantena. Paperin oma varovainen muotoilu on, että MIRA oli kaikkien kahdeksan sairauden osalta ”johdonmukaisesti lääkäreiden tasolla ja usein heidän yläpuolellaan”.

Myös jatkopäätökset kestivät arvioinnin: ne olivat suurelta osin hoitosuositusten mukaisia, lääkityksen kannalta turvallisia ja sairaalaanoton suhteen asianmukaisia. Kirjoittajat raportoivat, ettei viidessä turvallisuusluokassa (lääkeyhteisvaikutukset, munuaistoimintaan perustuva annostus, allergiat, QT-riski ja opioidien määrääminen) ollut yhtään vakavaa lääkitysvirhettä, ja että reseptit olivat oikein 467 tapauksessa 468:sta — samalla he huomauttavat, ettei järjestelmä ”saavuttanut 100% luotettavuutta”. Sellaisenaan tulos on vaikuttava: koko tapauksen läpi toimiva agentti, joka päätyi diagnoosissa lääkäreitä parempaan tulokseen.

Mutta voiton *muoto* on yhtä tärkeä kuin voitto itsessään. Paperia lukeneet riippumattomat asiantuntijat kiinnittivät huomiota siihen, mistä etumatka tosiasiassa tuli. Science Media Centren asiantun-

tijapaneelissa Sheffieldin yliopiston tohtori Wei Xing huomautti, että otsikkotulos — tekoäly voitti lääkärit diagnostisessa tarkkuudessa — johtui **enimmäkseen sairauksista, joissa tutkimustulos oli selkeä**, kuten umpilisäketulehduksesta ja haimatulehduksesta, joissa ratkaiseva kuvaus tai laboratorioarvo ratkaisee kysymyksen. Keuhkokuumeessa ja virtsatieinfektioissa — kahdessa hyvin tavallisessa päivystykseen hakeutumisen syyssä — *sekä* tekoäly että lääkärit pärjäsivät heikoimmin, ja niiden välinen ero oli pienin.

Osapuolten pelitavassa oli myös epäsymmetria, joka näkyy paperin omissa luvuissa. MIRA nojasi laboratorioon voimakkaammin: se käytti noin **51%**:a tavanomaisessa hoidossa saatavilla olevista laboratorioanalyyteistä, kun erikoislääkärit käyttivät noin **28%**:a — mediaanina seitsemän veriparametria enemmän tapausta kohti. Kirjoittajat korostavat, että tämä jäi aineiston oman tavanomaisen hoidon lähtötason *alapuolelle*, eli kyse ei ollut kaikenkattavasta testausstrategiasta, eikä kalliimpaa poikkileikkauskuvantamista käytetty järjestelmällisesti enemmän. Silti suurempi tietomäärä voi jo itsessään nostaa diagnostista tarkkuutta. Kyse ei siis aivan ole samanlaisilla tiedoilla toimivien osapuolten vertailusta, minkä Xing nosti suoraan esiin. Myös kliinikko näyttäisi paperilla terävämmältä, jos hän voisi määrätä kaikki halvat verikokeet pohtimatta kustannuksia, epämukavuutta tai viivettä.

Miksi ”päihitti lääkärit” ei tarkoita ”parempi lääkäri”

Vertailuvoitto on helppo tulkita liian pitkälle. ”MIRA sai paremman pistemäärän” ja ”MIRA on parempi” väliin mahtuu ainakin kolme asiaa.

Ensinnäkin **mihin tuloksia verrattiin**. Edinburghin yliopiston professori Julie Jacko huomautti, että useat MIRA:n keskeiset tulosmittarit on määritelty suhteessa siihen, mitä alkuperäiseen aineistoon oli dokumentoitu — eli järjestelmää palkitaan **kirjatun kliinisen toiminnan toistamisesta**, ei välttämättä optimaalisen hoidon osoittamisesta. Historiallisesta potilaskertomuksesta tulee vastausavain. Se on järkevä tapa rakentaa vertailu, mutta se mittaa yhdenmukaisuutta tehdyn hoidon kanssa, ei oikeellisuutta absoluuttisessa mielessä. Reiluuden vuoksi on todettava, että tämä ongelma osuu voimakkaimmin *hoidon yhdenmukaisuuden* mittareihin — vastasivatko MIRA:n määräykset potilaskertomusta? — kun taas otsikon diagnostinen tarkkuus pisteytetään kotiutusdiagnoosia vasten, mikä on lähempänä todellista lopputulosta kuin pelkkä dokumentaation toistaminen.

Toiseksi on jo mainittu **tiedon epäsymmetria**: paljon enemmän verikokeita — noin 51% saatavilla olevista analyyteistä verrattuna 28%:iin — tarkoittaa enemmän näyttöä. Osa tarkkuuserosta on todennäköisesti *ostettu* lisätiedolla, ei päätelty paremmin.

Kolmanneksi **ympäristö, jossa järjestelmä toimi**. Kyse oli hiekkalaatikosta, retrospektiivisistä tapauksista, kahdeksasta ennalta valitusta diagnoosista ja pelkästä tekstistä. Oxfordin yliopiston tohtori Dominic Oliver muistutti, että todellinen kliininen arvio riippuu paitsi siitä, mitä potilaat sanovat, myös siitä, miten he sen sanovat — sekä fyysisestä tutkimuksesta, havaitusta käyttäytymisestä ja kehonkielestä, joista tekstipohjainen, valmista potilaskertomusta lukeva agentti ei näe mitään. Ja potilas, jonka ongelma ei kuulu kahdeksan valitun diagnoosin joukkoon, on yksinkertaisesti tämän tutkimuksen ulottumattomissa.

Arvioijat nostivat esiin myös hiljaisemman huolen: **aineiston kontaminaation**. MIMIC-IV on

julkinen ja siitä on kirjoitettu paljon, joten avoimella internetillä koulutettu kielimalli on voinut jo nähdä artikkeleita, tapauskeskusteluja tai itse aineistoa. Sheffieldin yliopiston tohtori Midhun Parakkal Unni nosti tämän suoraan esiin — jos osa vastauksista oli mukana koulutusdatassa, osa suorituskyvystä voi olla muistamista eikä kliinistä päättelyä, eikä näitä voi erottaa ilman riippumatonta toisintoa. Kirjoittajat eivät sivuuta ongelmaa: he kirjoittavat, että tuloksia ”voitaisiin varovaisesti tulkita mahdolliseksi ylärajaksi” ja että ne ”saattavat yliarvioida yleistymistä muihin julkisiin tapauksiin” — paperi siis asettaa itse katon omalle otsikkotulokselleen.

Mikään tästä ei tee MIRA:sta vähemmän kiinnostavaa. Se tekee oikeasta tulkinnasta kalibroidumman: retrospektiivisessä vertailussa koko potilaskertomuksen läpi toimiva agentti päihitti lääkärit diagnooseissa, joissa tutkimustulokset olivat selkeitä — osittain tilaamalla enemmän tutkimuksia ja osittain olemalla samaa mieltä kirjatun potilaskertomuksen kanssa. Se on todellinen kyvykkyyden osoitus, ei tuomio siitä, että koneet diagnosoisivat nyt lääkäreitä paremmin.

Mitä tämä *ei* osoita

- Se **ei** osoita, että MIRA toimii oikeilla potilailla. Kaikki tapaukset olivat retrospektiivisiä ja simuloituja; se ei hoitanut yhtäkään todellista potilasta.
- Se **ei** osoita, että järjestelmä toimii kahdeksan diagnoosin ulkopuolella. Ennalta valitun joukon ulkopuolisia sairauksia — lääketieteen sotkuista ja erilaistumatonta enemmistöä — ei testattu.
- Se **ei** osoita, että järjestelmä olisi turvallinen ottaa käyttöön. Kirjoittajat itse toteavat, että yleistettävyyttä, turvallisuutta ja hallintaa on vielä tutkittava prospektiivisesti todellisessa ympäristössä.
- Se **ei** osoita reilua suoraa vertailua lääkäreihin. MIRA käytti paljon enemmän verikokeita (noin 51% saatavilla olevista analyyteistä verrattuna 28%:iin), ja osa hoitotuloksista pisteytettiin osittain kirjatun potilaskertomuksen eikä parhaan mahdollisen hoidon todellisen vertailuarvon mukaan.
- Se **ei** osoita, ettei tulos voisi sisältää muistamista. Julkinen MIMIC-IV-aineisto voi olla päällekkäinen mallin koulutusdatan kanssa, minkä poissulkeminen vaatisi riippumatonta toisintoa.
- Se **ei** tarkoita ”tekoäly korvaa lääkärit”. Kyse on päätöksenteon tuesta, joka voi *toimia* potilaskertomuksen läpi; arvioijien yhteinen näkemys on, että todellisessa käytössä se toimisi yhdessä kliinikkojen kanssa, joilla säilyy päätösvalta ja jotka tuovat mukaan kaiken sen, mikä jää tekstimuotoisen potilaskertomuksen ulkopuolelle.

Kuinka vahvaa näyttö on?

Väite kannattaa jakaa kahteen osaan, koska niiden näyttö on hyvin erilaista.

Kyvykkyyden osoituksena — että kielimalliin perustuva agentti pystyy hoitamaan koko kliinisen tapauksen potilastietojärjestelmän hiekkalaatikossa yhdistämällä anamneesin, tutkimukset, diagnoosin ja määräykset alusta loppuun — työ on aidosti uusi ja varsin vakuuttava. Juuri tämä osa on uutta, ja siihen kannattaa kiinnittää huomiota.

Ylivertaisuusväitteenä — että MIRA olisi *lääkäreitä parempi* — näyttö on rajattua ja sitä pitää

lukea varoen. Tulos pätee tiettyyn retrospektiiviseen vertailuun, kahdeksaan diagnoosiin, tiedon epäsymmetriaan (enemmän tutkimuksia) ja historiallisen potilaskertomuksen muodostamaan vastausavaimeen, ja mukana on todellinen mahdollisuus koulutusdatan kontaminaatiosta. Se riittää sanomaan: ”agentti suoriutui tässä vertailussa vaikuttavasti.” Se ei riitä sanomaan: ”agentti on lääkäriä parempi diagnosoija”, eivätkä kirjoittajat väitä jälkimmäistä.

Hyödyllisin asenne ei ole ”tekoäly voittaa lääkärit” eikä ”vain lelu”. Se on tämä: uudenlainen lääketieteellinen tekoäly — sellainen, joka *toimii* koko potilaskertomuksen läpi kysymyksiin vastaamisen sijaan — pärjäsi hyvin vaikeassa retrospektiivisessä vertailussa ja joutuu nyt osoittamaan kykynsä siellä, missä sitä ei ole vielä kokeiltu: oikeilla, ennalta luokittelemattomilla potilailla, prospektiivisesti ja valvotun hallinnan alla.

Miksi tämä on tärkeää

Viime vuosina lääketieteellisen tekoälyn kiinnostava kysymys on hiljalleen muuttunut. Ennen kysyttiin ”saako malli diagnoosin oikein?” — ja mallit vastasivat yhä uudelleen kyllä yhä siistimmissä testiaineistoissa. MIRA siirtää painopisteen vaikeampaan kysymykseen: ”pystyykö malli *tekemään työn* — keräämään tietoa, määräämään tutkimuksia, tulkitsemaan, päättämään ja toimimaan — koko työnkulun läpi?” Se on hyödyllisempi ja rehellisempi kysymys, koska juuri toiminnassa ovat sekä vaikeus että riski.

Siksi kehystys on tärkeä. Otsikkorefleksi — *tekoäly päihittää lääkärit* — osoittaa tuloksen vähiten uuteen ja heikoimmin tuettuun osaan. Aidosti uusi asia on pienempi mutta seurauksiltaan merkittävämpi: agentti, joka pystyy etenemään koko potilaskertomuksen läpi. Jos tämä kyvykkyys kestää jatkotestit, se muuttaa **työnkulkua** kauan ennen kuin se muuttaa sitä, kuka työnkulkua johtaa. Lääkäriä ei korvata; juuri tutkimusten määrääminen, tulosten tulkitseminen ja niiden perässä pysyminen — työnkulku — on ensimmäinen paikka, johon tällainen työkalu asettuu.

Samalla todistustaakka asettuu oikeaan suuntaan. Retrospektiivisten tapausten tulostauluvoitto on syy tehdä prospektiivinen tutkimus, ei sen korvike. Kirjoittajat sanovat tämän itse. Rehellinen tulkinta MIRA:sta on kutsu seuraavaan tutkimukseen — sillä standardilla, jonka ala jo tuntee: mitataan potilailla, ei vertailuaineistoilla.

Tiivistelmä

MIRA on autonominen tekoälyagentti, joka toimii hiekkalaatikkoon rajatussa sähköisessä potilastietojärjestelmässä: se voi ottaa anamneesin, määrätä ja tulkita laboratorio-, kuvantamis- ja mikrobiologisia tutkimuksia, päätyä diagnoosiin ja kirjoittaa hoitosuunnitelmia. Julkisen MIMIC-IV-tietokannan 574 retrospektiivisessä tapauksessa ja kahdeksassa ennalta valitussa diagnoosissa se päihitti lääkärit diagnostisessa tarkkuudessa (88.9% kokonaisuudessaan; 87.8% verrattuna 78.1%:iin suorassa vertailussa) ja teki suurelta osin hoitosuositusten mukaisia ja lääkitysturvallisia päätöksiä. Arviointi oli kuitenkin menneisiin potilastietoihin perustuva tekstipohjainen simulaatio; suuri osa

etumatkasta tuli sairauksista, joissa tutkimustulokset olivat selkeitä; MIRA käytti paljon enemmän verikokeita kuin lääkärit (noin 51% saatavilla olevista analyyteistä verrattuna 28%:iin); useita hoitotuloksia pisteytettiin sen mukaan, mitä alkuperäiseen potilaskertomukseen oli kirjattu; ja julkinen aineisto aiheuttaa todellisen koulutusdatan kontaminaatioriskin — jota kirjoittajat itse kutsuvat lukujensa mahdolliseksi ylärajaksi. Todellinen edistysaskel on agentti, joka toimii koko työnkulun läpi eikä vain vastaa yksittäisiin kysymyksiin. Kirjoittajat ovat selväsanaisia siitä, että yleistettävyys, turvallisuus ja hallinta vaativat edelleen prospektiivisia tutkimuksia todellisessa ympäristössä. Se on oikea tulkinta: vaikuttava kyvykkyyden osoitus, ei todiste siitä, että tekoäly diagnosoisi lääkäreitä paremmin eikä järjestelmä, joka olisi valmis oikealle klinikalle.

Realistinen arvio

Mitä paperi osoittaa: Kielimalliin perustuva agentti (MIRA) voi käydä koko kliinisen tapauksen läpi hiekkalaatikkoon rajatussa potilastietojärjestelmässä — anamneesi, tutkimukset, diagnoosi, määräykset — ja päihitti 574 tapauksen retrospektiivisessä, kahdeksan diagnoosin vertailussa lääkärit diagnostisessa tarkkuudessa (88.9% kokonaisuudessaan; 87.8% verrattuna erikoislääkärien 78.1%:iin) ilman vakavia lääkitysvirheitä.

Mikä on uskottavaa mutta ei osoitettu: Että kyvykkyys tuottaa hyötyä oikeille, ennalta luokittelemattomille potilaille; että tarkkuusetu johtuu paremmasta päättelystä eikä suuremmasta tutkimusmäärästä, yhdenmukaisuudesta historiallisen potilaskertomuksen kanssa tai julkisen aineiston muistamisesta.

Mitä se ei osoita: Että MIRA toimii kahdeksan diagnoosinsa ulkopuolella; että se on turvallinen ottaa käyttöön; että se päihittää lääkärit reilussa vertailussa, jossa osapuolet saavat saman tiedon; että se korvaa kliinikot; että tuloksissa ei ole koulutusdatan kontaminaatiota.

Tärkeimmät rajoitukset: Retrospektiivinen, simuloitu ja tekstipohjainen arviointi; kahdeksan ennalta valittua diagnoosia; tiedon epäsymmetria (paljon enemmän verikokeita — noin 51% saatavilla olevista analyyteistä verrattuna 28%:iin, nimenomaan edullisissa verikokeissa eikä kuvantamisessa); hoitotuloksia pisteytettiin osittain historiallisen potilaskertomuksen mukaan; ja julkinen vertailuaineisto (MIMIC-IV), jonka kielimalli on voinut nähdä koulutuksessa — minkä kirjoittajat nostavat mahdolliseksi suorituskyvyn ylärajaksi.

Kuinka paljon luottamusta tavallisella lukijalla pitäisi olla? Paljon siihen, että MIRA on todellinen ja uudenlainen kyvykkyyden osoitus — agentti, joka *toimii* potilaskertomuksen läpi eikä ole vain chatbot. Vähän siihen, että se olisi *lääkäriä parempi diagnosoija*: tämä väite rajoittuu yhteen retrospektiiviseen vertailuun, ja sitä sekoittavat tutkimusten määrääminen, pisteytys historiallisen potilaskertomuksen mukaan ja mahdollinen kontaminaatio. Kirjoittajien oma johtopäätös on oikea: järjestelmää on tutkittava prospektiivisesti todellisessa ympäristössä ennen kuin tulos merkitsee mitään potilashoidolle.

Lähteet

Ferber et al., Towards autonomous medical artificial intelligence agents, Nature (2026)
https://doi.org/10.1038/s41586-026-10675-5

PubMed 42310457 — peer-reviewed abstract
https://pubmed.ncbi.nlm.nih.gov/42310457/

Science Media Centre — expert reaction to MIRA and AMIE (2026)
https://www.sciencemediacentre.org/expert-reaction-to-presentation-of-two-new-medical-ai-models-for-

News-Medical (2026) — illustrates the 'outperforms doctors' framing
https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EH
aspx