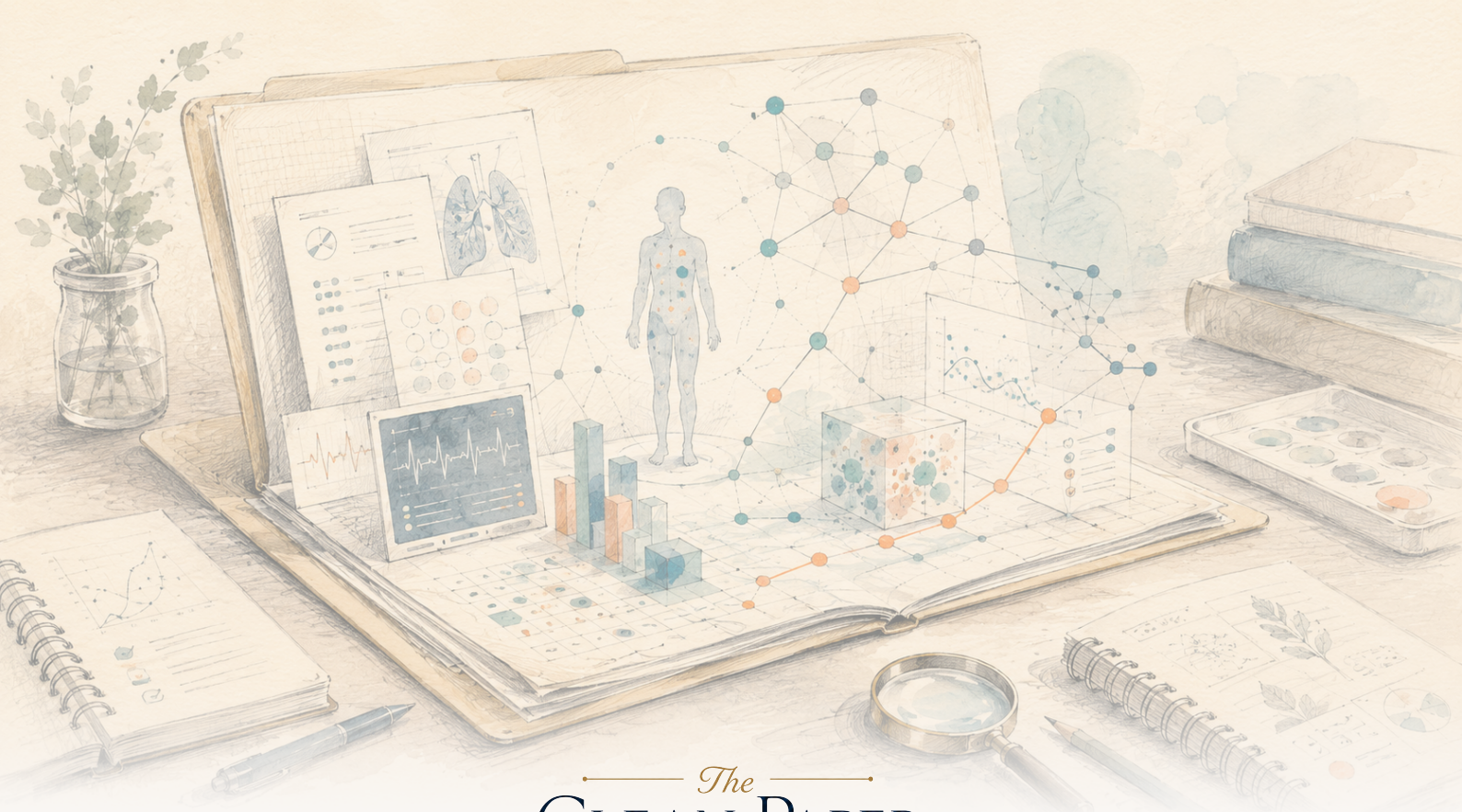


The Clean Paper

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

DI agentas savarankiškai tvarkė ištisus pacientų atvejus – simuliacijoje, su istoriniais įrašais

MIRA yra naujo tipo medicininis DI: užuot atsakęs į vieną klausimą, agentas izoliuotoje ligoninės įrašų simuliacijoje tvarko visą atvejį – renka anamnezę, užsako ir interpretuoja tyrimus, nustato diagnozę ir suformuoja nurodymus. 574 retrospektyviuose viešos duomenų bazės atvejuose, apimančiuose aštuonias iš anksto pasirinktas diagnozes, autoriai nurodo didesnį diagnostinį tikslumą nei gydytojų ir daugiausia gaires atitinkančius, vaistų požiūriu saugius sprendimus. Tačiau kiekviena šio sakinio išlyga svarbi: sistema veikė simuliacijoje su praeities įrašais ir vien tekstu; didelė persvaros dalis buvo ligose su aiškiais tyrimų rezultatais; ji naudojo maždaug dvigubai daugiau kraujo rodiklių nei gydytojai; keli rezultatai vertinti pagal tai, kas užfiksuota istoriniame įraše. Tikroji pažanga – agentas, galintis veikti per visą darbo eigą, o ne įrodymas, kad mašina diagnozuoja geriau už gydytoją. Patys autoriai pabrėžia, kad generalizacijai, saugai ir valdymui reikia prospektyvių realaus pasaulio tyrimų.

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

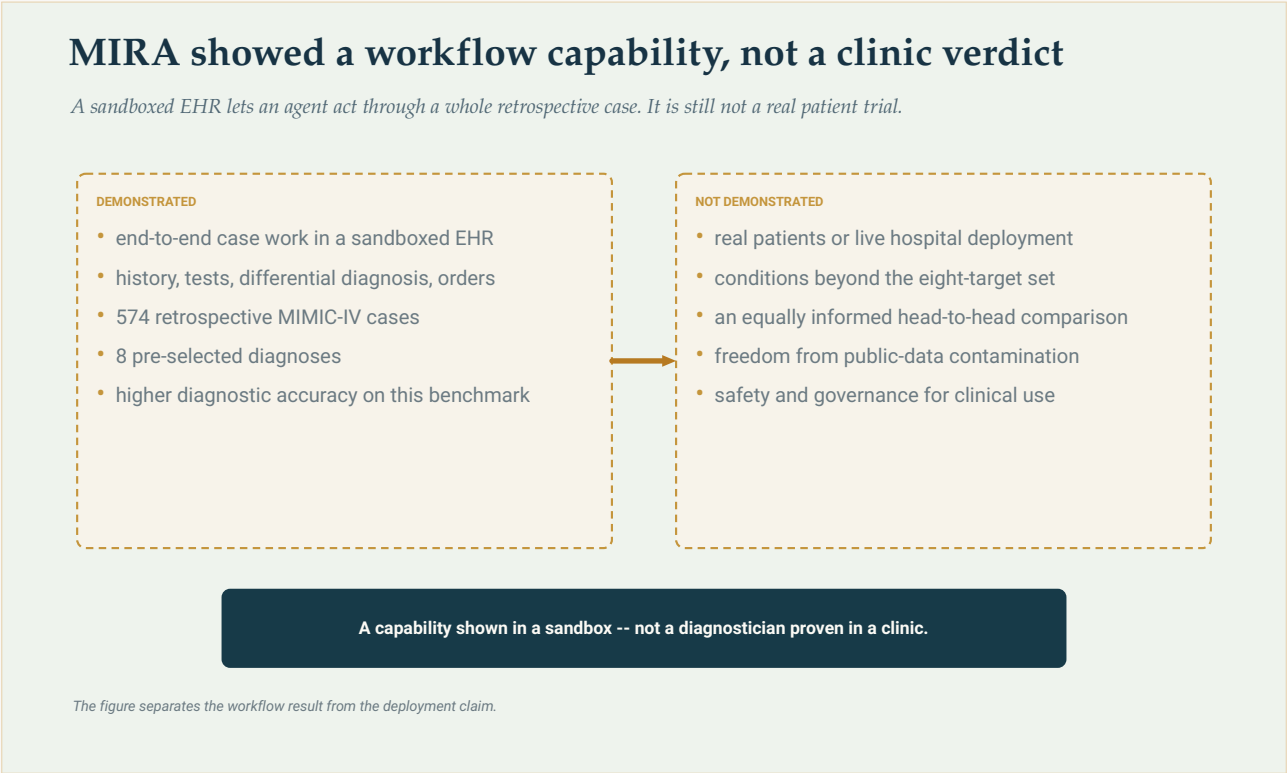
Atsakyti į klausimą – ne tas pats, kas savarankiškai tvarkyti visą atvejį

Dauguma istorijų apie medicininį DI yra apie modelį, kuris *atsako*. Jam pateikiama klinikinė situacija ar egzamino klausimas, jis grąžina diagnozę ar patarimų pastraipą ir gauna gerą įvertinimą. Naudinga, bet siaura: tarsi sumanus konsultantas, kuris niekada pats nedirba su paciento įrašu.

MIRA sukurta kitam darbui, ir būtent šis skirtumas yra tikroji naujiena. Užuot atsakiusi į vieną klausimą, sistema tvarko visą atvejį: skaito paciento įrašą, sprendžia, kokios anamnezės dar trūksta, užsako laboratorinius, vaizdinius ir mikrobiologinius tyrimus, perskaito jų rezultatus, siaurina diferencinę diagnozę ir galiausiai suformuoja tolesnius nurodymus – receptus, hospitalizavimą, siuntimą operacijai. Ji veikia *pačioje* elektroninėje sveikatos istorijoje, kaip gydytojas, o ne tik parašo tekstą, kurį žmogus turėtų perkelti ranka.

Tai tikras žingsnis nuo sistemos, kuri pataria, prie sistemos, kuri atlieka veiksmus per visą darbo eigą. Ir būtent toks žingsnis žiniasklaidoje lengvai susispaudžia iki „DI pranoko gydytojus“. Todėl svarbu tiksliai pasakyti, kas buvo išbandyta ir kokiomis sąlygomis.

Vieno sakinio versija: MIRA savarankiškai tvarkė ištisus atvejus ir šiame teste pagal diagnostinį tikslumą pranoko gydytojus, tačiau tai darė izoliuotoje simuliacijoje, su retrospektyviais įrašais, tik aštuonių iš anksto pasirinktų diagnozių ribose ir bendraudama vien tekstu; nemaža jos persvaros dalis atsirado ten, kur tyrimų rezultatai buvo ypač aiškūs. Pažanga tikra. Rezultatų lentelė nėra klinika.



MIRA pademonstravo gebėjimą nuo pradžios iki galo vykdyti klinikinę darbo eigą izoliuotoje elektroninės sveikatos istorijos aplinkoje. Tyrimas neparodė realaus diegimo klinikoje, plataus diagnostinio aprėpties diapazono ar visiškai lygiaverčio, vienodai informuoto palyginimo su gydytojais.

Original Aurora diagram — The Clean Paper CC BY 4.0

Ką padarė autoriai

MIRA – *Medical Intelligence for Reasoning and Action* – yra autonominis agentas, sukurtas naudojant OpenAI modelius: pokalbį ir veiksmus vykdanti dalis naudoja **GPT-4o**, o atskiram planavimo etapui taikomas OpenAI samprotavimo modelis **o1**. Visa tai veikia specialiai sukurtoje, standartus atitinkančioje sistemoje, paremtoje HL7 FHIR – tuo pačiu duomenų standartu, kuriuo realios ligoninės keičiasi medicinos įrašais. Agentui prieinama daugiau kaip aštuoniasdešimt tūkstančių galimų klinikinių veiksmų. Izoliuotoje aplinkoje MIRA gali gauti paciento anamnezę, užsakyti ir interpretuoti laboratorinius, vaizdinius bei mikrobiologinius tyrimus, sudaryti diferencinę diagnozę ir gydymo planą – skirti vaistus, planuoti operaciją ar hospitalizavimą. Viena svarbi detalė: automatiniai „teisėjai“, vertinę MIRA atsakymus, taip pat buvo GPT-4o – tos pačios modelių šeimos, kuri buvo testuojama. Po recenzento iškelto klausimo autoriai šį vertinimą papildė sertifikuoto gydytojo peržiūra.

Testų rinkinys paimtas iš **MIMIC-IV** – didelės, viešai prieinamos, nuasmenintos elektroninių sveikatos įrašų duomenų bazės. Iš maždaug 300 000 pacientų, gydytų Beth Israel Deaconess medicinos centre 2008–2019 m., autoriai atrinko **574 atvejus**, apimančius **aštuonias tikslines diagnozes** – pilvo organų patologijas ir skubias vidaus ligų būkles, tarp jų apendicitą, pankreatitą, pneumoniją ir šlapimo takų infekciją. Kiekvieną atvejį MIRA pradėdavo turėdama ribotą informaciją ir žingsnis po žingsnio turėdavo spręsti, ką daryti toliau. Darbo eiga priminė tikrą diagnostinį procesą, tik ji vyko su istoriniu įrašu, kurio galutinis rezultatas jau buvo žinomas.

Tada MIRA rezultatai tais pačiais atvejais buvo palyginti su gydytojų rezultatais.

Ką iš tikrųjų reiškia „autonominis agentas izoliuotoje EHR aplinkoje“?

Šie trys žodžiai neša daug reikšmės, todėl verta juos išskleisti.
Agentas reiškia, kad sistema nėra pokalbių robotas, vien atsakantis į užklausą. Ji veikia ciklu: įvertina dabartinę būseną, pasirenka veiksmą (pavyzdžiui, užsako tyrimą ar paklausia apie tam tikrą anamnezės detalę), gauna rezultatą ir pasirenka kitą veiksmą, kol prieina prie diagnozės ir plano. „Intelektą“ suteikia kalbos modelis, o *veikimo galimybę* – aplinkinė programinė sistema, kuri modelio tekstą paverčia leidžiamomis EHR operacijomis ir grąžina jam jų rezultatus.
Izoliuota EHR – tai simuliuota, nuo realios ligoninės atskirta medicinos įrašų sistemos kopija. MIRA gali „užsakyti“ tyrimą ir gauti rezultatą, kurį tas pacientas iš tikrųjų turėjo, nes atvejis istorinis ir atsakymas jau yra įraše. Nė vienas jos veiksmas nepaveikia tikro paciento ar tikros klinikos.
Autonominis reiškia, kad *simuliacijos viduje* sistema be žmogaus įsikišimo užbaigia atvejį nuo pradžios iki galo. Tai **nereiškia**, kad jai buvo leista be priežiūros gydyti pacientus. Tai du labai skirtingi teiginiai, ir tyrimas tikrino tik pirmąjį.
Dar viena svarbi simuliacijos ypatybė: MIRA gali *užsakyti* bet kokį pageidaujamą tyrimą, tačiau

> sistema gali pateikti rezultatą tik tada, jei tas tyrimas **iš tikrųjų buvo atliktas** tam pacientui realiame gyvenime. Jei pacientui buvo atliktas kraujo tyrimas, MIRA gauna tikruosius skaičius; jei ji paprašo vaizdinio tyrimo, kurio niekas tuo metu neužsakė, grąžinama „N/A – atlikti nepavyko“, o ne išgalvotas rezultatas. Tas pats galioja anamnezei: jei įraše nėra atsakymo į MIRA klausimą, simuliuotas pacientas tiesiog „nežino“. Taigi agentas negali išsikviesti lemiamo tyrimo, kuris realiame procese niekada nebuvo atliktas – jis gali naudotis tik tuo, ką iš tikrųjų apėmė istorinis ištyrimas.
> Šis skirtumas svarbus todėl, kad žodis „autonominis“ antraštėje lengviausiai suprantamas būtent klaidinga, platesne prasme.

Ką jie nustatė

Šiame teste **MIRA pagal diagnostinį tikslumą pranoko gydytojus**, tačiau po viena antrašte slepiasi trys skirtingi skaičiai, kuriuos verta atskirti.

Vertinant visus **574 atvejus**, MIRA teisingą diagnozę nurodė **88,9 %** atvejų; atsakymas buvo lyginamas su MIMIC-IV įraše fiksuota išrašymo diagnoze. Tiesioginiam palyginimui su žmonėmis gydytojai nesprendė visų 574 atvejų – jie dirbo su bendra **311 atvejų** imtimi ir turėjo tuos pačius įrašus bei įrankius kaip MIRA. Tais pačiais 311 atvejų MIRA pasiekė **87,8 %** tikslumą ir buvo lyginama su dviem atskirai surinktomis gydytojų grupėmis. Pirmoji buvo **vyresniųjų gydytojų grupė**: keturi sertifikuoti gydytojai, turintys 7–11 metų patirties, pasiekę **78,1 %**. Antroji – **mišrios patirties grupė**, daugiausia jaunesni rezidentai ir du specialistai, artimesnė realiam skubios pagalbos skyriaus personalo pasiskirstymui; ji pasiekė **71,1 %**. Tie patys atvejai, tie patys įrankiai – ir rezultatuose pirmavo DI, po jo sekė patyrę gydytojai, o treti buvo jaunesnių gydytojų dominuojama komanda. Pats straipsnis atsargiau formuluoja, kad pagal visas aštuonias ligas MIRA buvo „nuosekliai lygiavertė gydytojams ir dažnai juos pranoko“.

Tolesni sprendimai taip pat atrodė geri: dauguma jų atitiko gaires, vaistų saugos reikalavimus ir buvo tinkami hospitalizavimo požiūriu. Autoriai nurodo, kad penkiose vaistų saugos kategorijose (sąveikos, dozavimas pagal inkstų funkciją, alergijos, QT rizika ir opioidų skyrimas) nebuvo didelio sunkumo klaidų, o 467 iš 468 receptų buvo įvertinti kaip teisingi. Kartu jie aiškiai pažymi, kad sistema „nepasiekė 100 % patikimumo“. Pažvelgus vien į šiuos skaičius rezultatas įspūdingas: agentas tvarko visą atvejį ir diagnozės tikslumu lenkia gydytojus.

Tačiau svarbi ne tik pergalė, bet ir jos *forma*. Nepriklausomi specialistai, komentavę tyrimą, atkreipė dėmesį, iš kur daugiausia atsirado MIRA persvara. Science Media Centre ekspertų apžvalgoje dr. Wei Xing (Sheffieldo universitetas) pažymėjo, kad garsiausias rezultatas – DI persvara pagal diagnostinį tikslumą – buvo **daugiausia nulemtas ligų su aiškiais tyrimų rezultatais**, pavyzdžiui, apendicito ir pankreatito, kur lemiamas vaizdinis tyrimas ar laboratorinė reikšmė stipriai susiaurina atsakymą. Pneumonijos ir šlapimo takų infekcijų – dviejų labai dažnų skubios pagalbos priežasčių – atvejais *ir* DI, *ir* gydytojai pasirodė prasčiausiai, o skirtumas tarp jų buvo mažiausias.

Be to, abi pusės naudojosi informacija nevienodai. Tai matyti pačiame straipsnyje: MIRA laboratorija

rėmėsi intensyviau ir panaudojo apie **51 %** įprastoje priežiūroje prieinamų laboratorinių analitų, o sertifikuoti gydytojai – apie **28 %**; medianinis skirtumas buvo septyniais papildomais kraujo rodikliais vienam atvejui. Autoriai pabrėžia, kad tai vis tiek mažiau nei įprastinės priežiūros lygis pačiame duomenų rinkinyje ir kad MIRA sistemingai neužsakė daugiau brangių skerspjūvio vaizdinių tyrimų. Vis dėlto daugiau informacijos savaime gali didinti diagnostinį tikslumą, todėl tai nėra visai simetriškas vienodai informuotų dalyvių palyginimas. Xing šį skirtumą įvardijo tiesiogiai. Gydytojas, galintis laisvai užsakyti kiekvieną pigų kraujo tyrimą nesvarstydamas kainos, paciento diskomforto ar delsimo, popieriuje irgi atrodytų tikslesnis.

Kodėl „pranoko gydytojus“ nereiškia „yra geresnis gydytojas“

Geras rezultatas etalone lengvai perinterpretuojamas. Tarp „MIRA surinko daugiau taškų“ ir „MIRA yra geresnė“ stovi bent trys kliūtys.

Pirma – **su kuo buvo lyginamas atsakymas**. Profesorė Julie Jacko (Edinburgo universitetas) pažymėjo, kad keli svarbūs MIRA rodikliai apibrėžti pagal tai, kas buvo užfiksuota pradiniame duomenų rinkinyje. Kitaip tariant, sistema apdovanojama už **istoriniame įraše užfiksuoto klinikinio elgesio atkartojimą**, o ne būtinai už objektyviai geriausią priežiūrą. Istorinis paciento įrašas tampa atsakymų raktu. Tai pagrįstas būdas konstruoti etaloną, tačiau jis matuoja atitikimą tam, kas buvo padaryta, o ne absoliučią medicininę teisingumo reikšmę. Sąžiningumo dėlei, ši problema labiausiai liečia *gydymo atitikties* rodiklius – ar MIRA nurodymai sutapo su įrašu; pagrindinis diagnostinio tikslumo rodiklis lyginamas su išrašymo diagnoze, kuri yra arčiau realaus klinikinio rezultato nei vien dokumentacijos atkartojimas.

Antra – jau minėta **informacijos asimetrija**: daugiau kraujo tyrimų (apie 51 % prieinamų analitų prieš 28 %) reiškia daugiau įrodymų. Tikėtina, kad dalis tikslumo persvaros buvo ne vien „išmąstyta“, bet ir *nupirkta papildoma informacija*.

Trečia – **kur sistema veikė**. Tai buvo simuliacija su retrospektyviais atvejais, aštuoniomis iš anksto pasirinktomis diagnozėmis ir vien tekstine sąveika. Kaip pažymėjo dr. Dominic Oliver (Oksfordo universitetas), tikras klinikinis vertinimas priklauso ne tik nuo to, *ką* pacientas sako, bet ir *kaip* jis tai sako, taip pat nuo fizinio ištyrimo, elgesio ir kūno kalbos. Nieko iš to agentas, skaitantis užbaigtą tekstinį įrašą, nemato. O pacientas, kurio problema nepatenka į vieną iš aštuonių pasirinktų diagnozių, paprasčiausiai yra už šio tyrimo ribų.

Recenzentai iškėlė ir tylesnį rūpestį – **mokymo duomenų persidengimą**. MIMIC-IV yra vieša ir plačiai analizuojama bazė, todėl iš atvirojo interneto mokytas kalbos modelis galėjo būti matęs publikacijas, atvejų aptarimus ar net pačius duomenis. Dr. Midhun Parakkal Unni (Sheffieldo universitetas) tai įvardijo tiesiogiai: jei dalis atsakymų buvo mokymo duomenyse, dalis rezultato gali būti atmintis, o ne klinikinis samprotavimas; tai atskirti galėtų tik nepriklausomas pakartojimas. Svarbu, kad autoriai šios problemos nenumoja ranka: jie rašo, jog rezultatus „atsargiai galima vertinti kaip galimą viršutinę ribą“ ir kad jie „gali pervertinti generalizaciją į kitus viešus atvejus“. Straipsnis pats uždeda lubas savo garsiausiam skaičiui.

Nė viena iš šių išlygų MIRA nepadaro mažiau įdomios. Jos tik padeda rezultatą skaityti teisingu masteliu: retrospektyviame etalone agentas, galėjęs veikti per visą įrašą, diagnostiniu tikslumu pranoko gydytojus, ypač ligų su aiškiais tyrimais atvejais – iš dalies užsakydamas daugiau tyrimų, iš dalies sutapdamas su istoriniu įrašu. Tai tikras gebėjimų demonstravimas, bet ne nuosprendis, kad mašinos jau diagnozuoja geriau už gydytojus.

Ko tai *neįrodo*

- Tyrimas **neparodo**, kad MIRA veikia su tikrais pacientais. Visi atvejai buvo retrospektyvūs ir simuliuoti; nė vienas pacientas nebuvo jos gydomas.
- Jis **neparodo**, kad MIRA veikia už aštuonių diagnozių ribų. Didžioji netvarkinga ir iš anksto neatskirta medicinos dalis nebuvo testuota.
- Jis **neįrodo**, kad sistemą saugu diegti. Patys autoriai rašo, kad generalizacijai, saugai ir valdymui vis dar reikia prospektyvių realaus pasaulio tyrimų.
- Jis **nesukuria** visiškai lygiaverčio palyginimo su gydytojais. MIRA naudojo gerokai daugiau kraujo tyrimų (apie 51 % prieinamų analitų prieš 28 %), o keli gydymo rodikliai iš dalies buvo vertinami pagal istoriniame įraše užfiksuotus veiksmus, o ne nepriklausomą geriausios priežiūros etaloną.
- Jis **neparodo**, kad rezultatuose nėra įsiminimo poveikio. Vieši MIMIC-IV duomenys galėjo persidengti su modelio mokymo medžiaga; tam atmesti reikėtų nepriklausomo pakartojimo.
- Tai **nereiškia**, kad „DI pakeičia gydytojus“. Tai sprendimų palaikymo sistema, galinti atlikti veiksmus įraše; recenzentų vertinimu, realus jos naudojimas turėtų vykti kartu su klinicistais, kurie išlaiko sprendimo teisę ir pateikia tai, ko tekstinis įrašas negali perteikti.

Kiek tvirti yra įrodymai?

Teiginį verta padalyti į dvi dalis, nes jų įrodymų stiprumas labai skiriasi.

Kaip gebėjimų demonstravimas – kad kalbos modelio agentas gali izoliuotoje EHR aplinkoje nuo pradžios iki galo vykdyti klinikinį atvejį, nuosekliai jungdamas anamnezę, tyrimus, diagnozę ir nurodymus – darbas iš tiesų naujas ir gana įtikinamas. Būtent ši dalis yra naujovė ir verta dėmesio.

Kaip pranašumo prieš gydytojus teiginys įrodymai yra riboti ir juos reikia skaityti atsargiai. Persvara parodyta viename konkrečiame retrospektyviame etalone, aštuonių diagnozių ribose, esant informacijos asimetrijai (daugiau tyrimų), istoriniam įrašui atliekant dalies atsakymų rakto funkciją ir neišnykus galimam mokymo duomenų persidengimui. To pakanka teigti, kad „agentas šiame etalone pasirodė įspūdingai“. To nepakanka teigti, kad „agentas yra geresnis diagnostas už gydytoją“, ir patys autoriai antrojo teiginio nedaro.

Naudingiausia pozicija nėra nei „DI nugalėjo gydytojus“, nei „tai tik žaislas“. Tikslesnė išvada tokia: naujo tipo medicininis DI – sistema, kuri *veikia* visame įraše, o ne tik atsako į klausimą – gerai pasirodė sudėtingame retrospektyviame teste. Dabar ji turi įrodyti save ten, kur dar nebuvo išbandyta: prospektyviai, su tikrais ir iš anksto neatskirtais pacientais, esant realiai klinikinei

priežiūrai ir valdymui.

Kodėl tai svarbu

Per pastaruosius kelerius metus svarbiausias medicininio DI klausimas tyliai pasikeitė. Anksčiau klausta: „ar modelis gali nustatyti teisingą diagnozę?“ – ir vis švaresniuose testų rinkiniuose modeliai vis dažniau atsakydavo „taip“. MIRA žymi perėjimą prie sunkesnio klausimo: „ar modelis gali *atlikti darbą* – surinkti informaciją, užsakyti tyrimus, interpretuoti, nuspręsti ir veikti per visą darbo eigą?“ Tai naudingesnis ir sąžiningesnis klausimas, nes būtent veiksmuose kartu gyvena ir sudėtingumas, ir rizika.

Todėl svarbu, kaip rezultatas pateikiamas. Refleksinė antraštė *„DI pranoko gydytojus“* nukreipia dėmesį į mažiausiai naują ir silpniausiai apibendrinamą rezultato dalį. Tikroji naujovė kuklesnė, bet reikšmingesnė: agentas gali judėti per visą paciento įrašą. Jei šis gebėjimas išliks realiomis sąlygomis, jis pirmiausia keis **darbo eigą**, o tik daug vėliau – klausimą, kas jai vadovauja. Gydytojas neišnyksta; pirmiausia keičiasi užsakymų, interpretavimo ir rezultatų sekimo darbas.

Tai taip pat teisingai perkelia įrodinėjimo naštą. Pergalė retrospektyviame reitinge yra priežastis atlikti prospektyvų klinikinį tyrimą, o ne jo pakaitalas. Autoriai patys tai pripažįsta. Sąžiningiausias MIRA skaitymas – kvietimas atlikti kitą tyrimą pagal medicinai jau pažįstamą standartą: vertinti pacientus, o ne vien etalonus.

Trumpa santrauka

MIRA yra autonominis DI agentas, veikiantis izoliuotoje elektroninės sveikatos istorijos aplinkoje: jis gali rinkti anamnezę, užsakyti ir interpretuoti laboratorinius, vaizdinius bei mikrobiologinius tyrimus, nustatyti diagnozę ir sudaryti gydymo planus. Išbandyta su 574 retrospektyviais viešos MIMIC-IV bazės atvejais ir aštuoniomis iš anksto pasirinktoms diagnozėmis, MIRA pagal diagnostinį tikslumą pranoko gydytojus (88,9 % visoje imtyje; 87,8 % prieš 78,1 % tiesioginiame palyginime) ir priėmė daugiausia gaires atitinkančius, vaistų požiūriu saugius sprendimus. Tačiau tai buvo simuliacija su istoriniais įrašais ir vien tekstine sąveika; nemaža persvaros dalis atsirado ligų su labai aiškiais tyrimų rezultatais atvejais; MIRA naudojo gerokai daugiau kraujo tyrimų nei gydytojai (apie 51 % prieinamų analitų prieš 28 %); keli gydymo rodikliai buvo vertinami pagal tai, kas užfiksuota pradiniame įraše; o vieša duomenų bazė kelia realią mokymo duomenų persidengimo riziką, kurią patys autoriai įvardija kaip galimą viršutinę rezultatų ribą. Tikroji pažanga – ne atskiro klausimo atsakymas, o agentas, galintis veikti per visą darbo eigą. Autoriai aiškiai teigia, kad generalizacijai, saugai ir valdymui dar reikia prospektyvių realaus pasaulio tyrimų. Tai įspūdingas gebėjimų demonstravimas, o ne įrodymas, kad DI diagnozuoja geriau už gydytoją, ir ne sistema, parengta realiai klinikai.

Be pagražinimų

Ką rodo tyrimas: kalbos modelio agentas MIRA izoliuotoje EHR aplinkoje gali nuo pradžios iki galo vykdyti klinikinį atvejį – anamnezę, tyrimus, diagnozę ir nurodymus – ir 574 retrospektyvių atvejų, apimančių aštuonias diagnozes, teste diagnostiniu tikslumu pranoko gydytojus (88,9 % visoje imtyje; 87,8 % prieš 78,1 % sertifikuotų gydytojų grupėje), nepadarydamas didelio sunkumo vaistų klaidų.

Kas tikėtina, bet neįrodyta: kad šis gebėjimas duotų naudos realiems, iš anksto neatskirtiems pacientams; kad tikslumo persvara atspindi geresnį samprotavimą, o ne daugiau užsakytų tyrimų, sutapimą su istoriniu įrašu ar viešų duomenų įsiminimą.

Ko tyrimas nerodo: kad MIRA veikia už aštuonių diagnozių ribų; kad ją saugu diegti; kad ji pranoksta gydytojus visiškai sąžiningame ir vienodai informuotame palyginime; kad gali pakeisti klinicistus; kad rezultatai tikrai nepaveikti mokymo duomenų persidengimo.

Pagrindiniai ribotumai: retrospektyvus, simuliuotas, vien tekstinis vertinimas; aštuonios iš anksto pasirinktos diagnozės; informacijos asimetrija (gerokai daugiau kraujo tyrimų – apie 51 % prieinamų analitų prieš 28 %, daugiausia pigių laboratorinių, o ne vaizdinių tyrimų); gydymo rezultatų dalinis vertinimas pagal istorinį įrašą; viešas MIMIC-IV etalonas, kurį modelis galėjo būti matęs mokymo metu – ir patys autoriai tai įvardija kaip galimą viršutinę veikimo ribą.

Kiek pasitikėti turėtų bendrasis skaitytojas? Tvirtai – kad MIRA yra tikras ir naujas gebėjimų demonstravimas: agentas, kuris *veikia* paciento įraše, o ne vien kalbasi. Kur kas mažiau – teiginiu, kad ji yra *geresnė diagnostė už gydytoją*: ši išvada apsiriboja vienu retrospektyviu etalonu ir yra supainiota su skirtingu tyrimų užsakymu, vertinimu pagal įrašą ir galimu duomenų persidengimu. Pačių autorių galutinė išvada yra tinkama – prieš kalbant apie pacientų priežiūrą reikia prospektyvaus realaus pasaulio tyrimo.

Šaltiniai

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