



The CLEAN PAPER

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

AI エージェントが患者症例全体を自力で処理した——ただしシミュレーター内で、過去の記録を使って

MIRA は新しい種類の医療 AI だ。一つの質問に答えるのではなく、模擬病院記録の中で症例全体を進める。病歴を取り、検査を指示して結果を読み、診断に至り、指示まで作成する。公開データベースの過去 574 症例、事前に選ばれた 8 診断で、著者らは診断精度が医師を上回り、判断の多くがガイドラインに沿い、薬剤安全性も高かったと報告した。ただし、この一文の但し書きはすべて重要だ。実験は過去記録を使ったサンドボックス内のテキストのみで行われ、優位性のかなりの部分は検査結果が明瞭な疾患で生じ、MIRA は医師の約 2 倍の血液検査項目を使い、複数のアウトカムは元カルテに記録された内容を基準に採点された。本当の前進は、ワークフロー全体で行動するエージェントであることだ。機械が今や医師より優れた診断をする証拠ではない。著者ら自身、一般化、安全性、ガバナンスには前向きな実臨床研究がなお必要だと述べている。

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

質問に答えることと、症例全体を進めることは同じではない

医療 AI の話題の多くは、モデルが何かに答えるというものだ。症例要約や試験問題を与えると、診断名や助言の文章を返し、高い得点を取る。役には立つが、範囲は狭い。カルテには一切触れない、頭の切れるコンサルタントのようなものだ。

MIRA は、それとは違うことをするために作られている。そして、その違いこそがこの研究の本当のニュースだ。一つの質問に答えるのではなく、症例全体を進める。患者記録を読み、まだ必要な病歴情報を判断し、検査室検査、画像検査、微生物検査を指示し、結果を読み、鑑別診断を絞り込み、その後に必要な指示——処方、入院、外科への紹介——まで作成する。人間が転記するための自由文を出すのではなく、臨床医と同じように電子健康記録（EHR）の中で操作を行う。

これは確かに一歩前進だ。助言するシステムから、ワークフロー全体で行動するシステムへの移行である。同時に、報道ではまさにこうした前進が「AI が医師を上回った」という一言に平板化されやすい。だから、何がどこで試されたのかを正確に見る価値がある。

一文で言えばこうだ。MIRA は症例全体を自律的に処理し、このベンチマークでは診断精度で医師を上回った。ただし、それはサンドボックス内で、過去の記録を用い、あらかじめ選ばれた 8 つの診断を対象に、テキストだけでやり取りして行われた。また、優位性のかなりの部分は、検査結果が明瞭な疾患で得られていた。前進そのものは本物だ。ただし、ベンチマークの順位がそのまま診療能力を意味するわけではない。

MIRA showed a workflow capability, not a clinic verdict

A sandboxed EHR lets an agent act through a whole retrospective case. It is still not a real patient trial.

DEMONSTRATED

- end-to-end case work in a sandboxed EHR
- history, tests, differential diagnosis, orders
- 574 retrospective MIMIC-IV cases
- 8 pre-selected diagnoses
- higher diagnostic accuracy on this benchmark

NOT DEMONSTRATED

- real patients or live hospital deployment
- conditions beyond the eight-target set
- an equally informed head-to-head comparison
- freedom from public-data contamination
- safety and governance for clinical use

A capability shown in a sandbox -- not a diagnostician proven in a clinic.

The figure separates the workflow result from the deployment claim.

MIRA が実証したのは、EHR サンドボックス内で症例を最初から最後まで処理できる能力である。実臨床への導入、幅広い診断領域での性能、あるいは同じ情報を与えた医師との公平な直接比較を実証したわけでは

ない。

Original Aurora diagram —The Clean Paper CC BY 4.0

著者らが行ったこと

MIRA—Medical Intelligence for Reasoning and Action—は、OpenAI のモデルを用いて構築された自律型エージェントだ。会話と操作を担う部分には **GPT-4o**、別の計画ステップには OpenAI の推論モデル **o1** を使う。これらは、実際の病院が医療記録をやり取りするために使うデータ標準 HL7 FHIR に基づいた、標準準拠の独自フレームワークの中に置かれ、8 万種類を超える臨床操作のツール群を利用できる。そのサンドボックス内で MIRA は、患者の病歴を取得し、検査室検査、画像検査、微生物検査を指示・解釈し、鑑別診断を作り、治療計画を立てることができる。薬を処方し、手術を予定し、入院計画まで作る。最初に押さえておきたい細部が一つある。MIRA の回答を自動採点した「判定役」も GPT-4o、つまり評価対象と同じモデル系統だった。査読者がこの点を問題にした後、著者らは専門医資格を持つ医師によるレビューを追加して補強している。

テストセットは、匿名化された大規模な公開 EHR データベース **MIMIC-IV** から作られた。Beth Israel Deaconess Medical Center で 2008～2019 年に治療を受けた約 30 万人の患者記録から、著者らは **8 つの対象診断**にまたがる **574 症例**を選んだ。虫垂炎、膵炎、肺炎、尿路感染症など、腹部疾患や内科救急が含まれる。各症例で MIRA は限られた初期情報から始め、次に何をすべきかを一段ずつ判断した。形としては実際の診療プロセスに似ているが、使っているのは結末がすでに記録されている過去の症例だ。

その性能を、同じ症例に取り組んだ医師と比較した。

「EHR サンドボックス内の自律型エージェント」とは実際に何を意味するのか

ここでは三つの言葉が大きな意味を担っているの、分けて見ておこう。

エージェント (agent) とは、このシステムがプロンプトに答えるだけのチャットボットではないという意味だ。現在の状態を見る → 操作を選ぶ (この検査を指示する、この病歴を尋ねる) → 結果を見る → 次の操作を選ぶ、というループを、診断と計画に到達するまで繰り返す。言語モデルが推論を担い、その出力を許可された EHR 操作へ変換し、結果を再びモデルへ返す周囲の仕組みが自律的な行動を可能にする。

EHR サンドボックス (sandboxed EHR) とは、実際に稼働中の病院システムではなく、隔離された医療記録システムのシミュレーション環境を意味する。症例は過去のものなので、MIRA が検査を「指示」すると、その患者が実際に受けた検査結果を受け取れる。答えはすでに記録されているからだ。MIRA の操作が本物の患者や実際の診療に触れることはない。

自律型 (autonomous) とは、人間を途中で挟まずに症例を最初から最後まで完了するという意味だ——サンドボックスの中では。患者を監督なしで任せられたという意味ではない。この二つはまったく異なる主張で、試されたのは前者だけだ。

もう一つ、サンドボックスについて誤解しやすい点がある。MIRA は好きな検査を指示できるが、システムが結果を返せるのは、その患者が実際の診療でその検査を**本当に受けていた**場合だけだ。患者が受けた血液検査を依頼すれば実測値が返るが、誰も撮っていない画像検

査を頼めば「N/A — could not be performed」と返るだけで、架空の数値が生成されることはない。病歴も同じで、記録に MIRA が尋ねた情報がなければ、模擬患者は単に「分からない」と答える。したがって MIRA は、現実には行われなかった決定的な検査を都合よく呼び出すことはできない。実際の診療過程に含まれていた情報だけで仕事をする。この区別が重要なのは、見出しで最も目を引く「自律型」という言葉が、後者の意味に読み替えられやすいからだ。

分かったこと

このベンチマークでは、**MIRA は診断精度で医師を上回った**。ただし、見出しの下には別々の三つの数字があり、混同しやすいので分けておく必要がある。

MIRA 単独では、**574 症例**すべてを通じて **88.9%** で正しい診断名を挙げた。正解は MIMIC-IV で実際に記録された退院時診断を基準に採点された。人間との直接比較では、医師は 574 例すべてを担当したわけではなく、共通の **311 症例**に取り組み、MIRA と同じ記録とツールを使った。この同じ 311 例で MIRA は **87.8%** だった。比較対象となった医師は別々に集められた二群である。一つ目の**上級群**は、経験 7~11 年の専門医資格を持つ医師 4 人で、**78.1%**。二つ目は**経験年数混合群**で、主に若手研修医に専門医 2 人を加えた、実際の救急部門の人員構成に近いグループで、**71.1%** だった。同じ症例、同じツールで、順位は AI、経験豊富な医師、若手中心のチームとなった。論文自身の慎重な表現では、8 疾患すべてを通じて MIRA は医師と「一貫して同等で、しばしば上回った」。

その後の意思決定もおおむね良好だった。ガイドラインへの整合性が高く、薬物療法の安全性も高く、入院判断も適切だったと著者らは報告している。薬物相互作用、腎機能に応じた用量、アレルギー、QT リスク、オピオイド処方という五つの安全性カテゴリーで重篤な薬剤関連エラーはなく、処方 468 例中 467 例で正しかった。ただし著者ら自身、「100% の信頼性には達しなかった」と明記している。字面どおりに受け取れば、これは印象的な結果だ。症例全体を動かすエージェントが、診断で医師より高得点を取ったのだから。

しかし、勝ったという事実と同じくらい、どのように勝ったかが重要だ。論文を読んだ独立専門家は、優位性が実際にどこから来たのかを指摘している。Science Media Centre の専門家コメントで、Wei Xing 博士 (University of Sheffield) は、AI が診断精度で医師を上回ったという見出しの数字は、虫垂炎や膀胱炎のように決定的な画像や検査値で答えがはっきりする、**明瞭な検査結果を持つ疾患に主として支えられていた**と述べた。肺炎と尿路感染症 — 実際に救急部を受診する非常に一般的な理由の二つ — では、AI も医師も成績が最も悪く、両者の差も最も小さかった。

また、両者の「戦い方」には非対称性があり、そのことは論文自身の数字に表れている。MIRA は検査室データをより多く使った。通常診療で利用可能だった検査項目の約 **51%** を参照したのに対し、専門医資格を持つ医師は約 **28%** で、1 症例あたり中央値で 7 項目多い血液パラメータを使っていた。著者らは、これは何でも検査する戦略ではなく、データセット内の通常診療そのもののベースラインより少ないと慎重に位置付け、高額になりやすい断層画像検査が系統的に増えたわけではないとも報告している。それでも、情報量が増えるだけで診断精度は上がり得る。したがって、同じだけ情報を持った者同士の完全な同条件比較とは言えず、Xing もこの差を明確に指摘している。費用、不快感、遅延を考えず安価な血液検査を好きなだけ指示できる臨床医も、紙の上ではより鋭

く見えるだろう。

「医師に勝った」は「より優れた医師」と同じではない

ベンチマークの勝利は過大解釈しやすい。「MIRA の得点が高かった」と「MIRA のほうが優れている」の間には、少なくとも三つの問題がある。

第一は、**何を正解として採点したか**だ。Julie Jacko 教授 (University of Edinburgh) は、MIRA の主要アウトカムのいくつかが元データセットに記録されていた内容との一致で定義されていると指摘した。つまり、このシステムは必ずしも最適な医療を示すことではなく、**記録された臨床行動を再現すること**で報われる。過去のカルテが解答集になるわけだ。ベンチマークを作る方法としては合理的だが、測っているのは絶対的な意味での正しさではなく、実際に行われたこととの一致である。論文に公平を期せば、この問題が最も強く効くのは治療整合性の指標——MIRA の指示がカルテと一致したか——である。一方、見出しとなった診断精度は退院時診断を基準にしており、単なる記録の再現よりは実際の結果に近い。

第二は、すでに触れた**情報量の非対称性**だ。利用可能な検査項目の約 51% 対 28% という大きな差は、そのまま証拠の量の差でもある。精度差の一部は、推論そのものより、追加の情報を買った結果である可能性が高い。

第三は、**どこで動かしたか**である。評価は EHR サンドボックス内で過去症例を使い、あらかじめ選ばれた 8 診断だけを対象に、テキストのみで行われた。Dominic Oliver 博士 (University of Oxford) が述べるように、実際の臨床評価は患者が何を言うかだけでなく、どのように言うかにも依存する。身体診察、観察される行動、ボディランゲージも含まれるが、完成した記録を読むテキスト専用エージェントにはそのどれも見えない。そして、患者の問題が選ばれた 8 診断のどれにも当てはまらない場合、この研究からは何も言えない。

査読者が挙げた、もう少し静かな懸念もある。**データ汚染**だ。MIMIC-IV は公開され、多数の論文や解説で扱われているため、公開インターネット上のデータで学習した言語モデルが、その論文、症例解説、あるいはデータ自体をすでに見ていた可能性がある。Midhun Parakkal Unni 博士 (University of Sheffield) はこの点を直接指摘した。もし答えの一部が訓練データに含まれていたなら、性能の一部は臨床推論ではなく記憶によるものであり、両者を分けるには独立した追試が必要になる。重要なのは、著者らがこの問題をごまかしていないことだ。結果は「可能な上限値として慎重に解釈できる」「他の公開症例への一般化を過大評価している可能性がある」と自ら書いている。論文自身が、その派手な数字に上限線を引いている。

だからといって MIRA が面白くなるわけではない。必要なのは読み方の調整だ。過去症例のベンチマークでは、記録全体を操作できるエージェントが、検査で答えが明瞭な診断において医師を上回った。その一部は検査をより多く頼んだためであり、一部は記録されたカルテと一致したためでもある。これは確かな能力実証だが、機械が医師より優れた診断をするようになったという判決ではない。

この研究からは言えないこと

- MIRA が実際の患者で機能することは示していない。すべて過去症例のシミュレーションで、MIRA が本物の患者を診療したことは一度もない。
- 8つを超える診断で機能することは示していない。事前選択された集合の外にある疾患 — 医学の大半を占める、整理されていない複雑な症例 — は試されていない。
- 導入して安全であることは示していない。著者ら自身、一般化、安全性、ガバナンスには前向きな実臨床研究がなお必要だと書いている。
- 医師との公平な直接比較を確立したわけではない。MIRA ははるかに多くの血液検査（利用可能項目の約 51% 対 28%）を使い、治療アウトカムのひとつかは真に最善の医療ではなく、記録されたカルテとの一致によって部分的に採点された。
- 結果が記憶の影響を受けていないことは示していない。公開 MIMIC-IV データがモデルの学習データと重なっている可能性があり、排除には独立した再現研究が必要だ。
- 「AI が医師を置き換える」という意味ではない。これは記録全体で操作できる意思決定支援であり、査読コメントの共通した見方は、現実の利用は臨床医との協働になるというものだ。権限は臨床医が持ち、テキスト記録から抜け落ちる情報も人間が補う。

証拠をどう評価するか

主張を二つに分けるべきだ。両者では証拠の強さがまったく違う。

能力実証として — 言語モデルのエージェントが EHR サンドボックス内で、病歴、検査、診断、指示をつなぎながら、臨床症例全体を端から端まで処理できるという点 — この研究は本当に新しく、かなり説得力がある。ここが新規性のある部分であり、注目すべきところだ。

優越性の主張として — MIRA が医師より優れているという点 — 証拠には明確な境界があり、慎重に読む必要がある。特定の過去症例ベンチマーク、8 診断、情報の非対称性（検査数の多さ）、過去のカルテから作られた解答集、さらに訓練データ汚染の現実的な可能性という条件の下で成り立つ。「このベンチマークでエージェントは印象的な性能を示した」と言うには十分だ。「このエージェントは医師より優れた診断者だ」と言うには足りず、著者らも後者を主張していない。

最も有用な立場は、「AI が医師に勝った」でも「ただのおもちゃ」でもない。より正確にはこうだ。質問に答えるのではなく記録全体で行動する新しい種類の医療 AI が、難しい過去症例ベンチマークで良好な成績を示した。次に証明すべきなのは、まだ試されていない場所 — 実際の、診断の絞られていない患者を対象に、前向きに、適切なガバナンスの下で — である。

なぜ重要なのか

ここ数年、医療 AI で本当に面白い問いは静かに変わってきた。以前は「モデルは正しい診断を出せるか」だった。そして、ますます整えられたテストセットでは、モデルは何度も「できる」と答えてきた。MIRA は、より難しい問いへの移行を示している。「モデルはワークフロー全体で、情報を

集め、検査を指示し、解釈し、決め、行動する——つまり仕事そのものをできるか」。これはより有用で、より誠実な問いだ。なぜなら、難しさもリスクも、実際に行動するところにあるからだ。

だからこそ、結果の枠付けが重要になる。AI が医師を上回ったという見出し反射は、この結果の中で最も新規性が低く、最も裏付けの弱い部分を指している。本当に新しいものは、より小さく、しかしより大きな意味を持つ。記録全体を移動しながら操作できるエージェントだ。この能力が現実でも持ちこたえるなら、「誰が指揮するか」が変わるずっと前に、**ワークフロー**が変わる。医師が置き換わるのではない。検査を頼む、結果を解釈する、結果を追う——そうしたワークフローこそ、この種のツールが最初に入り込む場所だ。

そして、証明責任も正しい方向へ戻る。過去症例のランキングで勝ったことは、前向き試験を行う理由にはなるが、その代わりにはならない。著者らもそう述べている。MIRA を誠実に読むなら、これは次の研究への招待状だ。ベンチマークではなく患者で測るという、この分野がすでに知っている基準に従って。

まとめ

MIRA は、EHR サンドボックスを操作する自律型 AI エージェントである。病歴を取り、検査室検査・画像・微生物検査を指示して解釈し、診断に至り、治療計画を書くことができる。公開 MIMIC-IV データベースの過去 574 症例、あらかじめ選ばれた 8 診断で試験され、診断精度で医師を上回り（全体 88.9%、直接比較で 87.8% 対 78.1%）、判断の多くはガイドラインに沿い、薬剤安全性も高かった。ただし評価は過去記録を用いたテキストのみのシミュレーションだった。優位性のかなりの部分は検査結果が明瞭な疾患で生じ、MIRA は医師よりはるかに多くの血液検査項目を利用した（利用可能項目の約 51% 対 28%）。治療アウトカムのいくつかは元カルテに記録された内容との一致で採点され、公開データセットには学習データ汚染という現実的なリスクもある。著者ら自身、その性能値は上限値の可能性があると述べている。本当の前進は、孤立した質問に答えるのではなく、ワークフロー全体で行動できるエージェントであることだ。著者らは、一般化、安全性、ガバナンスには前向きな実臨床研究が必要だと明言している。したがって正しい読み方は、印象的な能力実証ではあるが、AI が医師より優れた診断をする証拠でも、実際の診療に投入できるシステムでもない、というものだ。

冷静に見ると

論文が示したこと：言語モデルのエージェント（MIRA）は、EHR サンドボックス内で、病歴、検査、診断、指示まで臨床症例全体を最初から最後まで進めることができた。8 診断・574 症例の過去データベンチマークでは、重篤な薬剤関連エラーなしに、診断精度で医師を上回った（全体 88.9%、専門医資格を持つ医師との比較で 87.8% 対 78.1%）。

もっともらしいが、まだ証明されていないこと：この能力が実際の、診断が整理されていない患者にも利益をもたらすこと。精度の差が、より多くの検査、記録カルテとの一致、公開データの記憶ではなく、より優れた推論そのものを反映していること。

示していないこと：MIRA が 8 診断の外でも機能すること。安全に導入できること。同じ情報を与えた公平な比較で医師を上回ること。臨床医を置き換えられること。結果に学習データ汚染がないこと。

主な限界：過去データを使ったシミュレーション、テキストのみの評価。事前選択された 8 診断。情報量の非対称性（低コストの血液検査項目をはるかに多く利用 — 約 51% 対 28%。画像検査ではない）。治療アウトカムの一部が過去カルテを基準に採点されていること。そして、言語モデルが学習時に見ていた可能性のある公開ベンチマーク MIMIC-IV を使っていること。著者ら自身、性能の上限値になっている可能性を指摘している。

一般読者はどの程度信頼すべきか？ MIRA が、単なるチャットボットではなく、記録全体で行動する新しい種類の能力実証であることには高い確信を持ってよい。一方、MIRA が医師より優れた診断者だという主張への確信は低い。この結論は一つの過去症例ベンチマークに限られ、検査指示量、カルテを基準とした採点、データ汚染の可能性に左右される。著者ら自身の結論が適切だ。患者診療に意味を持つと言う前に、前向きな実臨床研究が必要である。

出典

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>