

Künstliche Intelligenz macht Medizin: Ist es schon so weit?

Ein Microsoft-System löst schwere Lehrfälle besser als Ärzte. Ob das dem Patienten nützt, entscheidet sich auf einer Ebene, die die Werbung ausspart.

ZUSAMMENFASSUNG

Hintergrund. Ein Preprint von Microsoft löst schwere Fälle aus dem New England Journal of Medicine mit hoher Trefferquote und wird als Schritt zu einer überlegenen maschinellen Diagnostik dargestellt. Mehrere randomisierte Studien prüfen, was solche Sprachmodelle in der ärztlichen Diagnostik tatsächlich leisten.

Evidenz. Die Behauptung zerfällt in drei getrennt geprüfte Ebenen. Auf einem festen Satz von Fällen trifft das System die Diagnose mit hoher Rate. Ob ein Arzt mit dem System besser diagnostiziert, hängt an einer vorgeschalteten Schulung. Ob das Paar aus Arzt und System sicher bleibt, wenn das System falsch liegt, ist am schlechtesten belegt.

Offen. Ob die Werte aus ausgewählten Lehrfällen in unausgelesenen Praxispopulationen Bestand haben, ist nicht gezeigt. Die Zurechnung der Haftung bei schwer erklärbaren Systemen bleibt anzuwenden.

Schlussfolgerung. Die Evidenz ist umgekehrt zur Werbung verteilt. Sie steht am stärksten dort, wo das System wirbt, und am schwächsten dort, wo es für den Patienten zählt.

Dr. med. Christian Mauch · Facharzt für Orthopädie und Unfallchirurgie · Lesezeit 9 Minuten

Microsoft stellte im Sommer 2025 ein System vor, das mehrere Sprachmodelle wie ein Konsil zusammenschaltet, und meldete dazu eine Zahl. MAI-DxO löste 85,5 % einer Sammlung besonders schwerer Fälle aus dem New England Journal of Medicine, Ärzte ohne Hilfsmittel rund 20 %.¹ Aus solchen Zahlen wird der Schluss gezogen, künstliche Intelligenz diagnostiziere inzwischen besser als der Arzt. Der Schluss überspringt drei Fragen, die getrennt zu beantworten sind.

Ein System, das die ärztliche Diagnostik verbessern soll, müsste dreierlei leisten. Es müsste auf einem festen Satz von Fällen die richtige Diagnose treffen. Es müsste den Arzt, der es nutzt, besser machen als den Arzt ohne es. Und es müsste sicher bleiben, wenn es selbst falsch liegt, weil am Ende der Arzt für den Befund entsteht. In der öffentlichen Darstellung fließen die drei Leistungen zu einem Versprechen zusammen. In den Studien bleiben sie getrennt. Eine hohe Trefferquote auf einem Benchmark sagt nichts darüber, ob ein Arzt mit dem System besser wird, und noch weniger darüber, was geschieht, wenn das System irrt. Vermengt man die drei Ebenen, erscheint ein System als erwiesen, das auf der Ebene, die für den Patienten zählt, nie geprüft wurde.

Erste Ebene (KI allein): Stellt die KI die richtige Diagnose?

Auf dieser Ebene ist die Evidenz am stärksten. Das Microsoft-System arbeitete 304 klinisch-pathologische Konferenzfälle des New England Journal of Medicine aus den Jahren 2017 bis 2025 ab, ausgewählte Lehrfälle mit gesicherter Diagnose. In der besten Konfigura-

tion löste es 85,5 % davon, bei Kosten von rund 7.184 US-Dollar Rechenaufwand je Fall und etwa 11 Prozentpunkten über gängigen Einzelmodellen.¹ Die Trefferquote beschreibt, wie oft das System aus der Falldarstellung die gesicherte Diagnose ableitet.

Der Wert trägt eine Einschränkung, die in der Berichterstattung selten mitgenannt wird. Die Sammlung enthielt keine Gesunden und keine leichten Verläufe. Sie besteht aus schweren Rätseln, deren Auflösung feststeht. Der Wert misst die diagnostische Obergrenze an gelösten Aufgaben, nicht die Leistung in einer Praxis, in der Häufigkeit und Triage über den Verlauf entscheiden. Die Arbeit ist ein Preprint und nicht klinisch validiert. Über einen Arzt, der das System nutzt, sagt sie nichts, denn geprüft wurde das System allein. Das Fehlen leichter Fälle vergrößert zudem den scheinbaren Abstand zu den Ärzten, die im selben Test ohne Vorauswahl arbeiteten.

Zweite Ebene (Arzt mit KI): Wird der Arzt mit der KI besser?

Hier wird die Evidenz uneinheitlich, und sie hängt an einer Bedingung. Eine randomisierte Studie aus Stanford gab 50 Ärzten GPT-4 an die Hand, ohne vorherige Schulung. Sie bearbeiteten sechs komplexe Diagnosefälle nach Art einer Facharztprüfung, bewertet auf einem Raster für diagnostisches Denken bis 100 Punkte. Die Ärzte mit GPT-4 und die ohne kamen auf praktisch denselben Wert, 76 gegenüber 74 Punkten, ein Abstand ohne statistische Bedeutung. GPT-4 allein erreichte 92 %.² Der bloße Zugang zur KI verbesserte die Diagnose der Ärzte nicht.

Eine randomisierte Studie aus Pakistan verschob die Bedingung. 60 Ärzte durchliefen zuerst eine zwanzigstündige Schulung im Umgang mit Sprachmodellen, in deren Fähigkeiten, Grenzen und in der Prüfung ihrer Ausgaben. Danach wurden sie einem KI-Zugang oder der herkömmlichen Recherche zugewiesen und bearbeiteten bis zu sechs Fallvignetten. Die Ärzte mit Zugang erreichten im Mittel 71,4 % gegenüber 42,6 % ohne Zugang, ein bereinigter Abstand von 27,5 Prozentpunkten.³ In knapp einem Drittel der Fälle übertraf das Team aus Arzt und System das System allein. Die Arbeit prüfte Vignetten, keine realen Patienten, und die ausgewertete Gruppe umfasste 58 Ärzte. Was die Studie zeigt, ist eng umrissen. Erst die Schulung verwandelt den Zugang in einen Gewinn. Was mit einem fehlerhaften Vorschlag geschieht, prüfte sie nicht.

Dritte Ebene : Merkt der Arzt wenn die KI einen Fehler macht

Auf der Ebene, die über den Schaden am Patienten entscheidet, ist die Evidenz am dünnsten und am beunruhigendsten. Dieselbe Forschungsgruppe prüfte denselben geschulten Ärztetyp unter einer anderen Bedingung. 44 Ärzte, alle mit der zwanzigstündigen Schulung, wurden auf zwei Gruppen verteilt und erhielten sechs Vignetten. Die eine Gruppe erhielt fehlerfreie Vorschläge von ChatGPT-4o, die andere in drei der sechs Fälle absichtlich fehlerhafte. Die Ärzte durften die KI frei konsultieren und behielten die Entscheidung über jede Diagnose. Die Gruppe mit den fehlerhaften Vorschlägen fiel auf 73,3 % gegenüber 84,9 % in der Gruppe mit fehlerfreien Vorschlägen, ein bereinigter Abstand von 14,0 Prozentpunkten.⁴

Die Schulung hatte ausdrücklich die Prüfung der Ausgaben und das Erkennen von Halluzinationen zum Gegenstand. Sie verhinderte die Übernahme des Fehlers nicht. Der Effekt

trägt in der Literatur den Namen Automation Bias, die Neigung, einer maschinellen Ausgabe stärker zu vertrauen, als ihre Zuverlässigkeit es rechtfertigt. Der Befund benennt die Grenze der Schulung genau. Sie hebt die Leistung im Regelfall und senkt die Anfälligkeit im Fehlerfall nicht. Die Arbeit prüfte absichtlich eingebaute Fehler an Vignetten und misst damit nicht, wie häufig ein System im Alltag irrt. Sie zeigt, was geschieht, wenn es irrt.

Wie schmal die Basis dieser Systeme bleibt, zeigt eine Untersuchung an Kinderfällen. Eine ältere Version von ChatGPT bearbeitete hundert pädiatrische Fallrätsel und lag bei einer Fehlerquote von 83 %.⁵ Die Version des Systems und das Fachgebiet verschieben das Bild erheblich, und die hohe Trefferquote der ersten Ebene ist an ein bestimmtes System und einen bestimmten Fallkorpus gebunden.

3 Herangehensweisen unterschiedliche Gewichtungen

Für eine Werbeaussage nimmt man am liebsten die Zahl, die gut aussieht — hier die 85,5 % aus Ebene eins. Für die Frage, ob man ein System gefahrlos in der Praxis einsetzen kann, braucht man aber eine ganz andere Auskunft: was passiert, wenn es irrt. Und ausgerechnet dafür gibt es die dünnste Studienlage. Beweislage und klinische Bedeutung laufen also auseinander: Wo am meisten geprüft wurde, ist am wenigsten in Gefahr; wo am wenigsten geprüft wurde, steht der Patient. Die Schwelle zum Medizinprodukt

MAI-DxO und vergleichbare Systeme sind Forschungsarbeiten, keine zugelassenen Medizinprodukte. Eine Software, die eine Diagnose vorschlägt oder einen Befund erstellt, fällt nach ihrer Zweckbestimmung unter die Verordnung (EU) 2017/745 und braucht eine Konformitätsbewertung mit CE-Kennzeichnung und den Nachweis des klinischen Nutzens.⁶ Ein Preprint erfüllt das nicht. Die ärztliche Verantwortung entfällt durch den Einsatz eines Algorithmus nicht. Nach dem Behandlungsvertragsrecht schuldet der Arzt die Behandlung nach dem fachlichen Standard, und die Letztentscheidung bleibt bei ihm.⁷ Die Verantwortungsbereiche von Hersteller, Betreiber und Arzt bestehen bereits. Neu und schwierig ist ihre Anwendung auf Systeme, deren Vorschläge sich schwer nachvollziehen lassen. Die Studie zur dritten Ebene gibt dieser Frage einen konkreten Anlass. Sie zeigt, dass auch geschulte Prüfung die Übernahme des falschen Vorschlags nicht auffängt.

Ärztliche Einordnung

Die Befundlage lässt einen realistischen Entwicklungspfad erkennen. Die Systeme lösen ausgewählte schwere Aufgaben mit hoher Trefferquote. Ihr Nutzen für den Arzt entsteht erst nach einer Schulung im Umgang und dort, wo die Aufgabe ein Abwägen verlangt. Der sicherheitskritische Schritt, das Verhalten im Fehlerfall, ist am wenigsten untersucht und fällt in den vorliegenden Daten ungünstig aus. Daraus folgt keine Ablehnung und keine Euphorie. Es folgt eine Reihenfolge. Ein System gehört dort in die Praxis, wo das Zusammenspiel aus Arzt und System auch im Fehlerfall sicher bleibt, nicht schon dort, wo das System auf dem Benchmark gut abschneidet. Die Schulung ist notwendig, und sie genügt nicht. Sie muss mit einer Gestaltung einhergehen, die Unsicherheit des Systems sichtbar macht und in den entscheidenden Fällen eine unabhängige Prüfung erzwingt.

Die SummitGO-Perspektive

Der Fall zeigt ein Grundprinzip für medizinische KI insgesamt. Eine hohe Benchmark-Zahl entscheidet nicht über den klinischen Wert eines Systems. Über den Wert entscheidet, auf welcher Ebene die Zahl entstanden ist und ob die sicherheitskritische Ebene geprüft wurde. Die entscheidende Frage an ein System betrifft nicht seine Testleistung. Sie lautet, ob das Gesamtsystem aus Arzt und KI sicher versagt, wenn die KI falsch liegt. Die Leitlinie der assistierenden Intelligenz gewinnt damit einen empirischen Grund. Die geschulte Prüfung fängt den Fehler des Systems nicht auf, also bleibt die letzte Entscheidung beim Arzt. SummitGO liest die Studienlage als Herausgeber, der einordnet.

Fazit

Belegt ist, dass ein System auf ausgewählten schweren Fällen eine hohe diagnostische Trefferquote erreicht und dass geschulte Ärzte mit einem Sprachmodell im diagnostischen Denken besser werden. Belegt ist ebenso, dass dieselben geschulten Ärzte die Fehler des Systems übernehmen und an Boden verlieren, sobald die Vorschläge fehlerhaft sind. Offen ist, ob die Trefferquoten aus ausgewählten Lehrfällen in unausgelesenen Praxispopulationen Bestand haben und wie sich die Haftung bei schwer erklärbaren Systemen verteilt. Bis dahin ist der sichere Einsatz am Patienten eine Frage der Gestaltung und der erzwungenen Prüfung, nicht der Benchmark-Zahl.

QUELLEN

1. Nori H, et al. Sequential Diagnosis with Language Models (MAI-DxO). Microsoft Research, Preprint 2025. arXiv:2506.22405. arxiv.org/abs/2506.22405
2. Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Network Open 2024;7(10):e2440969. doi.org/10.1001/jamanetworkopen.2024.40969
3. Qazi IA, Ali A, Khawaja AU, et al. Large language model diagnostic assistance for physicians in a lower-middle-income country: a randomized controlled trial. Nature Health 2026;1:198–205. doi.org/10.1038/s44360-025-00007-8
4. Qazi IA, Ali A, Khawaja AU, et al. Automation Bias in Large Language Model-Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy: A Randomized Clinical Trial. NEJM AI 2026;3(5). doi.org/10.1056/AIoa2501001
5. Barile J, et al. Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies. JAMA Pediatrics 2024;178(3):313–315. doi.org/10.1001/jamapediatrics.2023.5750

6. Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates über Medizinprodukte (MDR), Art. 2, Anhang VIII Regel 11. ABl. EU L 117, 2017. eur-lex.europa.eu/eli/reg/2017/745/oj
7. Bürgerliches Gesetzbuch, § 630a Abs. 2. gesetze-im-internet.de

Eine erste, kürzere Fassung dieses Themas erschien im SummitGO-Newsletter „KI und Medizin“, Ausgabe 2.

HERAUSGEBER

Summitgo GmbH & Co. KG, Relenbergstraße 65, 70174 Stuttgart. Vertreten durch die Komplementärin Dr. Mauch Beteiligung GmbH (Amtsgericht Stuttgart HRB 805274), diese durch den Geschäftsführer Dr. Christian Mauch. Registergericht Amtsgericht Stuttgart, HRA 743417. info@summitgo.de. USt-IdNr. gemäß § 27a UStG: [folgt nach Erteilung]. Verantwortlich für den Inhalt nach § 18 Abs. 2 MStV: Dr. Christian Mauch, Anschrift wie vor.

TRANSPARENZ UND INTERESSENKONFLIKT

Herausgeber der Reihe ist die Summitgo GmbH & Co. KG, ein im Feld der medizinischen künstlichen Intelligenz tätiges Unternehmen. Der Beitrag behandelt Sprachmodelle in der Diagnostik, ein Feld, in dem SummitGO selbst tätig ist. Diese Interessenlage wird offengelegt. Auswahl, Bewertung und ärztliche Einordnung erfolgen durch den Autor auf Grundlage öffentlich zugänglicher Veröffentlichungen. Der Beitrag nennt kein Produkt und enthält keine Leistungsbewertung. Die tragenden Zahlen wurden gegen den Volltext der Primärquellen geprüft.