

Die künstliche Intelligenz hat schon einen Nobelpreis bekommen

Eine KI-Produktzulassung garantiert nicht den Patientennutzen. Woran ein KI-Versprechen für die Praxis prüfbar wird.

ZUSAMMENFASSUNG

Hintergrund. Der Chemie-Nobelpreis 2024 würdigte mit AlphaFold ein Verfahren des maschinellen Lernens. Seine Vorhersage weist zu jedem Ergebnis eine eigene Sicherheit aus und lässt sich an der experimentell bestimmten Struktur prüfen. In der klinischen Medizin fehlt ein vergleichbar unabhängiger und rasch verfügbarer Maßstab, weshalb dort die prospektive Prüfung an seine Stelle tritt.

Evidenz. Die FDA-Liste führt zum 30. Dezember 2025 insgesamt 1.451 freigegebene KI-Medizinprodukte, die Mehrzahl über das 510(k)-Verfahren, das die Gleichwertigkeit zu einem Vorgängerprodukt prüft, ohne den Nutzen am Patienten zu belegen. Eine Vollerhebung der 70 orthopädischen Produkte fand 8,6 % mit prospektiver klinischer Studie, bei 22,8 % lagen keine klinischen Daten vor. Von 21 geprüften Sprachmodellen lag die Fehlerquote bei der Differenzialdiagnose über 80 %, bei der Enddiagnose unter 40 %.

Offen. Ob die freigegebenen Produkte die Versorgung verbessern, ist für die überwiegende Zahl nicht belegt. Dafür fehlt die prospektive Prüfung am Patienten. Auch die ausgewiesene Unsicherheit je Einzelergebnis fehlt den meisten klinischen Systemen.

Schlussfolgerung. Ob ein KI-Versprechen für die Praxis trägt, hängt an der Anlage seiner Prüfung.

Dr. med. Christian Mauch · Facharzt für Orthopädie und Unfallchirurgie · Lesezeit 7 Minuten

2024 ging der Nobelpreis für Chemie zur Hälfte an David Baker für das rechnerische Entwerfen von Proteinen, die andere Hälfte gemeinsam an Demis Hassabis und John Jumper für die Vorhersage von Proteinstrukturen.¹ Ausgezeichnet wurde damit ein Verfahren des maschinellen Lernens, dessen Ergebnis sich unmittelbar überprüfen lässt. AlphaFold2 berechnet aus der Abfolge der Aminosäuren die dreidimensionale Faltung eines Proteins und gibt zu jeder Vorhersage ein Maß ihrer eigenen Sicherheit an.² Bis Oktober 2024 hatten mehr als zwei Millionen Menschen aus 190 Ländern das Modell genutzt. Die zugehörige Datenbank enthält rund 200 Millionen vorhergesagte Strukturen.³ Für die Sprechstunde hat dieses Modell keine Bedeutung. Für die

Beurteilung medizinischer KI hat es eine, weil es die Bedingung sichtbar macht, unter der eine Vorhersage überprüfbar wird.

Diese Bedingung ist ein unabhängiger, wohlfeiler und rasch verfügbarer Maßstab. Für ein Protein liefert ihn die experimentell aufgeklärte Struktur, gegen die sich jede Vorhersage einzeln stellen lässt. In der Medizin ist der entsprechende Maßstab teuer, verzögert und selbst fehlerbehaftet. Der klinische Verlauf zeigt sich Monate später, die histologische Sicherung kostet einen Eingriff, und die ärztliche Zweitbeurteilung streut. Wo der Maßstab schwer zu haben ist, verschiebt sich das Gewicht auf die Anlage der Prüfung. Die prospektive Untersuchung am Patienten ersetzt hier den Vergleichswert, den die Sache selbst nicht hergibt.

Was eine Marktfreigabe prüft

Die amerikanische Arzneimittelbehörde FDA führt eine öffentliche Liste der KI-gestützten Medizinprodukte, die in den Vereinigten Staaten in Verkehr gebracht werden dürfen. Sie weist zum Stand 30. Dezember 2025 insgesamt 1.451 Geräte aus. 295 kamen allein im Jahr 2025 hinzu, und 1.104, also etwa drei Viertel, entfallen auf die Radiologie.⁴ Diese Zahlen werden regelmäßig als Beleg für die klinische Reife des Feldes angeführt. Belegt ist mit ihnen die Zahl der Marktzugänge.

Der Weg dorthin führt in der Mehrzahl der Fälle über das 510(k)-Verfahren. Der Hersteller weist darin nach, dass sein Produkt einem bereits im Verkehr befindlichen Vorgängerprodukt im Wesentlichen gleichwertig ist. Geprüft wird die Gleichwertigkeit, nicht der Nutzen für den Patienten. Daraus folgt eine Eigenschaft des Verfahrens, die in der Debatte um Gerätezahlen selten benannt wird. Die Gleichwertigkeit bildet eine Kette. Jedes neue Produkt beruft sich auf ein Vorgängerprodukt, dieses wiederum auf ein früheres. Die klinische Evidenz, die am Anfang der Kette stand, wird mit keinem weiteren Glied erneuert. Je länger die Kette, desto weiter liegt die ursprüngliche Prüfung von dem Produkt entfernt, das heute im Behandlungszimmer steht.

In der Europäischen Union entscheidet über die Einstufung als Medizinprodukt die vom Hersteller festgelegte Zweckbestimmung. Die Bauart der Software ist dafür nicht maßgeblich. Trifft die Verordnung (EU) 2017/745 zu, durchläuft das Produkt ein Konformitätsbewertungsverfahren und trägt anschließend eine CE-Kennzeichnung.⁵ Diese bescheinigt Sicherheit und Leistungsfähigkeit für die erklärte Zweckbestimmung. Ein Nachweis, dass der Einsatz dem Patienten nützt, ist damit nicht verbunden. Das amerikanische und das europäische System beantworten dieselbe Frage, nämlich ob ein Produkt in Verkehr gebracht werden darf. Ob sein Einsatz die Versorgung verbessert, prüfen beide nicht.

Was aus freigegebenen Produkten wird, hat eine Auswertung von 903 in den USA freigegebenen KI-Produkten nachgezählt.⁶ 43 davon wurden

zurückgerufen, im Median 458 Tage nach der Freigabe. Bei Produkten ohne klinische Leistungsstudie lag die Rückrufquote bei 7,8 %, bei Produkten mit Studie bei 2,6 %. Die adjustierten Schätzer dieser Arbeit schließen die Eins ein und tragen die Aussage nicht. Für den Punkt, um den es hier geht, genügen die rohen Anteile. Der Rückruf trifft das Produkt, das ohne klinische Prüfung in den Markt gelangt ist, häufiger, und er trifft es früh.

Die Prüftiefe im eigenen Fach

Wie tief die Prüfung tatsächlich reicht, lässt sich an einem abgegrenzten Feld messen. Eine Arbeitsgruppe der Johns Hopkins University untersuchte im JAAOS Global Research and Reviews alle 70 KI-Medizinprodukte, die die FDA bis Februar 2025 für orthopädische Indikationen freigegeben hatte.⁷ Auf eine formale prospektive klinische Studie stützten sich 8,6 %. Bei 22,8 % lagen überhaupt keine klinischen Daten vor, 68,6 % wurden an rückblickenden Datensätzen geprüft. Bei den älteren Freigaben lag der Anteil höher: Von den zwischen 2017 und 2019 freigegebenen Produkten hatten 62,2 % keine klinische Prüfung, zwischen 2022 und 2024 waren es 19,7 %. Aussagekräftig an dieser Arbeit ist die Vollerhebung. Sie erfasst den gesamten freigegebenen Bestand eines Fachs. Ihre Schwäche liegt in der Quelle: Ausgewertet wurden die öffentlichen Zusammenfassungen der Behörde, nicht die vollständigen Studienunterlagen. Ein Produkt kann besser geprüft sein, als seine Zusammenfassung ausweist.

Die 8,6 % messen nicht den klinischen Nutzen der Produkte. Sie messen, was der Freigabeweg verlangt. Ein prospektiv geprüftes Produkt leistet deshalb nicht mehr. Es ist besser beurteilbar. Für die einzelne Praxis ist das die wichtigere Größe, weil sie über die Übertragbarkeit auf die eigene Population entscheidet. Der Anstieg der Freigaben von durchschnittlich 3,0 Geräten jährlich zwischen 2017 und 2019 auf 16,6 zwischen 2022 und 2024 ändert daran nichts.⁷ Die Menge wächst schneller als die Prüftiefe.

Wo eine hohe Trefferquote in die Irre führt

Der zweite Fehlschluss betrifft die Kennzahl selbst. Eine Querschnittsstudie im JAMA Network Open prüfte 21 allgemein verfügbare Sprachmodelle an 29 standardisierten Fallvignetten aus einem medizinischen Handbuch.⁸ Die Angaben wurden in der Reihenfolge der Sprechstunde zugespielt: erst die Anamnese, dann die körperliche Untersuchung, dann das Labor. Bei der Differenzialdiagnose lag die Fehlerquote in allen Modellen über 80 %, bei der Enddiagnose unter 40 %. Am Umgang mit Unsicherheit scheiterten die Modelle ebenso. Geprüft wurden die Grundmodelle mit abgeschalteter Reasoning-Einstellung, was die Übertragbarkeit auf die vermarkteten Fassungen begrenzt. Die Arbeitsgruppe um Arya Rao hält die Modelle für den unbeaufsichtigten klinischen Einsatz für nicht reif.⁸

Der Wert dieser Untersuchung liegt in der Trennung der Teilaufgaben. Gemessen wird üblicherweise der Endpunkt, an dem sämtliche Befunde vorliegen. Das ist der Punkt, an dem der Arzt die Antwort in aller Regel bereits kennt. Die ärztliche Leistung liegt davor: in der Auswahl der Fragen, im Aufstellen der konkurrierenden Möglichkeiten und in der Entscheidung, welche Untersuchung die Unsicherheit am stärksten senkt. An dieser Stelle brechen die Modelle ein. Eine Trefferquote, die am Endpunkt erhoben wird, beschreibt somit die leichteste Teilaufgabe des Verfahrens. Die Grenze der Arbeit ist ebenso deutlich zu benennen. Fallvignetten aus einem Lehrbuch sind vollständig, geordnet und frei von der Unschärfe einer realen Anamnese. Patienten waren nicht beteiligt. Über die Leistung an unausgewählten Patienten sagt die Arbeit nichts.

Ein drittes Merkmal von AlphaFold fehlt den meisten klinischen Systemen. Das Modell gibt zu jeder einzelnen Vorhersage aus, wie sicher es ist, und macht damit die schwachen Stellen der eigenen Ausgabe kenntlich.² Ein Befundsystem, das jeden Vorschlag mit gleicher Bestimmtheit anzeigt, verlagert diese Unterscheidung vollständig auf den Arzt, ohne ihm die Information dafür bereitzustellen. Wer ein Werkzeug beurteilt, fragt deshalb zuerst nach der ausgewiesenen Unsicherheit und erst danach nach der Trefferquote. Eine Angabe je Ergebnis ist der praktisch belastbarere Ausweis als eine Quote über den gesamten Testsatz.

Die rechtliche Schwelle

Wird eine Software außerhalb ihrer erklärten Zweckbestimmung zur Unterstützung einer konkreten Patientenentscheidung eingesetzt, kann sich die Verantwortung des Anwenders erheblich erhöhen. Ungeklärt ist die Haftung damit nicht. Sie richtet sich nach dem Einzelfall, nach der Zweckbestimmung des Produkts, nach der Fehlerursache, nach der Einbindung in den Behandlungsablauf und nach dem Verhalten von Hersteller, Betreiber und Arzt. Verantwortungsbereiche aller drei bestehen bereits.⁵ Neu und schwierig ist ihre Anwendung auf Modelle, deren Verhalten sich nicht vollständig erklären lässt. Praktisch bedeutsam bleiben die ärztliche Prüfung des Ergebnisses und deren Dokumentation.

Vor der Einführung eines Produkts sind daher fünf Angaben zu erheben. Erstens die Indikation und die Population, an der die genannte Kennzahl gemessen wurde. Zweitens die Anlage der Prüfung, prospektiv oder rückblickend. Drittens der Vergleichsmaßstab, gegen den gemessen wurde. Viertens die Angabe der Unsicherheit zu jeder Ausgabe des Systems. Fünftens die Herkunft der Zahl, unabhängige Prüfung oder Angabe des Herstellers. Fehlt eine dieser Angaben, lässt sich die Kennzahl auf die eigene Praxis nicht übertragen.

Ärztliche Einordnung

Die belastbare Aussage über medizinische KI folgt nicht aus der Zahl der freigegebenen Produkte und nicht aus der höchsten berichteten Trefferquote. Sie folgt aus der Anordnung, in der ein System geprüft wurde. Bei AlphaFold besteht diese Anordnung aus vier Elementen: einer scharf umrissenen Aufgabe, einem unabhängigen Maßstab, einer ausgewiesenen Unsicherheit zu jedem einzelnen Ergebnis und einer nachgeschalteten experimentellen Bestätigung. Bakers Arbeitsgruppe prüft die entworfenen Proteine im Labor. Die Vorhersage entscheidet nur, welcher Entwurf diese Prüfung lohnt.³ Fehlt eines dieser vier Elemente, verschiebt sich die Beweislast auf die klinische Studie. Die orthopädische Vollerhebung zeigt, dass diese Studie in der Regel nicht vorliegt.⁷ Die Arbeit von Rao und Kollegen zeigt, dass die veröffentlichte Kennzahl häufig die falsche Teilaufgabe misst.⁸ Beide Befunde führen auf dieselbe ärztliche Aufgabe: die Übersetzung einer Kennzahl in eine Aussage über die eigene Population, und die Feststellung, wo diese Übersetzung nicht gelingt.

Die SummitGO-Perspektive

Für SummitGO zeigt der Fall ein Grundprinzip medizinischer KI. Die Freigabe eines Produkts regelt den Marktzugang. Sie ist keine Aussage über den Nutzen im Behandlungsablauf und keine Entlastung des Arztes von der Prüfung des Einzelergebnisses. Ein System, das eine umrissene Aufgabe zuverlässig bearbeitet, kann in einem anderen Ablauf, an einer anderen Population und mit einer anderen Fragestellung unbrauchbar sein. Daraus folgt die Haltung des Herausgebers: Die Software bereitet vor, der Arzt entscheidet und verantwortet. Der Wert eines Werkzeugs bemisst sich an der Frage, welche Aufgabe es unter welchen Bedingungen zuverlässig stützt, und woran der Anwender das im eigenen Betrieb feststellen kann.

Fazit

Belegt ist, dass ein Modell des maschinellen Lernens ein wissenschaftliches Versprechen einlösen kann, wenn ein unabhängiger Maßstab die Vorhersage prüfbar macht. Belegt ist ferner, dass die Marktfreigabe eines KI-Medizinprodukts diesen Maßstab nicht enthält, dass in der Orthopädie 8,6 % der freigegebenen Produkte prospektiv geprüft sind und dass die höchsten Trefferquoten der Sprachmodelle an der Teilaufgabe erhoben werden, die klinisch am wenigsten fordert. Nicht belegt ist, dass die freigegebenen Produkte die Versorgung verbessern. Für diese Aussage fehlt in der überwiegenden Zahl der Fälle die prospektive Prüfung am Patienten.

QUELLEN

1. Pressemitteilung: The Nobel Prize in Chemistry 2024. Königlich-Schwedische Akademie der Wissenschaften; 2024. nobelprize.org/prizes/chemistry/2024/press-release
2. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589. doi:10.1038/s41586-021-03819-2
3. They have revealed proteins' secrets through computing and artificial intelligence. Populärwissenschaftlicher Hintergrund zum Nobelpreis für Chemie 2024. Nobelprize.org; 2024. nobelprize.org/prizes/chemistry/2024/popular-information
4. Artificial Intelligence-Enabled Medical Devices. AI-Enabled Medical Device List, Stand 30. Dezember 2025. U.S. Food and Drug Administration; 2026. fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
5. Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte. eur-lex.europa.eu/eli/reg/2017/745/oj
6. Ren Y, et al. Clinical Evidence and FDA Recalls of Artificial Intelligence-Enabled Medical Devices. JAMA Netw Open. 2026;9(6):e2617920. doi:10.1001/jamanetworkopen.2026.17920
7. Lee B, Jay M, Fox H, Padley J, Dai T, Levin AS. FDA-Cleared Artificial Intelligence Medical Devices in Orthopaedic Surgery. JAAOS Glob Res Rev. 2026;10(2):e25.00170. doi:10.5435/JAAOSGlobal-D-25-00170
8. Rao AS, Esmail KP, Lee RS, et al. Large Language Model Performance and Clinical Reasoning Tasks. JAMA Netw Open. 2026;9(4):e264003. doi:10.1001/jamanetworkopen.2026.4003

Eine erste, kürzere Fassung dieses Themas erschien im Newsletter „KI und Medizin“, Ausgabe 5.

TRANSPARENZ UND INTERESSENKONFLIKT

Herausgeber dieser Reihe ist die Summitgo GmbH & Co. KG, ein im Feld der medizinischen künstlichen Intelligenz tätiges Unternehmen. Der Beitrag behandelt die Prüfung und Freigabe medizinischer KI-Produkte, also ein Feld, in dem SummitGO selbst tätig ist. Diese wirtschaftliche Interessenlage wird offengelegt. Auswahl, Bewertung und ärztliche Einordnung erfolgen durch den Autor auf Grundlage öffentlich zugänglicher wissenschaftlicher Veröffentlichungen. Es handelt sich um einen redaktionellen Übersichtsbeitrag, nicht um Auftragsforschung und nicht um eine begutachtete Originalarbeit. Der Beitrag enthält keine Leistungsbehauptung zu einem Produkt der Summitgo GmbH & Co. KG. Eine externe Finanzierung bestand nicht.

IMPRESSUM

Summitgo GmbH & Co. KG, Relenbergstraße 65, 70174 Stuttgart. Komplementärin: Dr. Mauch Beteiligung GmbH (Amtsgericht Stuttgart HRB 805274), vertreten durch den Geschäftsführer Dr. Christian Mauch. Registergericht Amtsgericht Stuttgart, HRA 743417. Telefon +49 170 181 11 48, 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. © 2026 SummitGO.