TOPIC

KI in der Allergologie und Pneumologie

Diagnostik, Therapie und klinischer Alltag

Andreas Jerrentrup, Marburg

Nahezu jede Ärztin und jeder Arzt hat inzwischen praktische Erfahrungen mit künstlicher Intelligenz (KI) gesammelt, insbesondere mit Large-Language-Modellen (LLM) wie ChatGPT. Zum Zeitpunkt der Erstellung dieses Beitrags war ChatGPT 5.4 gerade erschienen und zeigte im medizinischen Kontext eine gegenüber früheren Versionen deutlich verbesserte Performance. Die Fortschritte dieser Modelle eröffnen für die medizinische Diagnostik und Therapie vielfältige Möglichkeiten der Unterstützung, sind aber zugleich durchaus auch mit verschiedenen Risiken verbunden.

Der Begriff **Künstliche Intelligenz** ist ein Oberbegriff für sehr unterschiedliche technologische Ansätze. Für eine sachgerechte Bewertung des klinischen Nutzens und der Grenzen dieser Verfahren ist es wesentlich, zwischen zwei derzeit besonders prägenden Entwicklungen zu unterscheiden: dem **klassischen maschinellen Lernen (ML)** und der **generativen KI**.

Klassisches maschinelles Lernen

Zu den etablierten KI-Verfahren gehört das klassische **maschinelle Lernen** (*Machine Learning*, ML), dessen besondere Stärke in der Erkennung komplexer Muster liegt. Solche Systeme beruhen häufig auf künstlichen neuronalen Netzen, insbesondere auf ***Convolutional Neural Networks*** (CNNs).

Das künstliche Neuron stellt dabei die elementare Recheneinheit dar. In Analogie zum biologischen Vorbild empfängt es gewichtete Eingangssignale, summiert diese und transformiert das Ergebnis mithilfe einer Aktivierungsfunktion in ein Ausgangssignal, das an die nächste Schicht weitergegeben wird. Funktionell lassen sich neuronale Netze daher als eine Kaskade aufeinander aufbauender Filter verstehen.

Das Training erfolgt in der Regel im Sinne eines **überwachten Lernens** (*Supervised Learning*). Hierzu wird das System mit großen Mengen annotierter Datensätze trainiert, beispielsweise mit tausenden Lungenfunktionskurven. Während dieses Prozesses zerlegt das Netz die Eingangsdaten in elementare Merkmale wie Kanten, Texturen oder Farb- bzw. Signalverteilungen. Mittels einer vom „gewünschten“ Zielergebnis rückwärts hin zum Eingangssignal gerichteten Iteration (*Backpropagation*) werden die Gewichtungen zwischen den künstlichen Neuronen schrittweise so angepasst, bis die Modellausgabe möglichst genau mit der durch Expertinnen und Experten vorgegebenen Referenzdiagnose übereinstimmt. Das System lernt somit durch wiederholte Fehlerkorrektur.

Nach Abschluss des Trainings kann ein solches Modell die erlernten Muster auf neue, bislang unbekannte Daten anwenden. In Vergleichsstudien zur Interpretation von **EKGs**, **Lungenfunktionen** oder **Mammografien** [3] erreichen entsprechende Algorithmen mittlerweile Sensitivitäten und Spezifitäten, die jene von Fachärztinnen und Fachärzten häufig übertreffen. Der wesentliche klinische Nutzen liegt dabei vor allem in der **Entlastung des ärztlichen Personals:** Repetitive Aufgaben der Mustererkennung können automatisiert werden, sodass mehr Zeit für klinische Einordnung, Therapieplanung und Patientenführung bleibt.

Die zentrale Limitation klassischer ML-Systeme besteht jedoch in ihrer ausgeprägten **Spezialisierung**. Es handelt sich gewissermaßen um hochperformante, aber eng fokussierte Einzwecksysteme. Eine KI, die auf die Befundung von Lungenfunktionsuntersuchungen trainiert wurde, ist nicht automatisch in der Lage, andere Pathologien zu erkennen, etwa einen Lungenrundherd in der Computertomografie.

Generative KI

„Wir behandeln Patienten, keine Laborwerte“: Dass medizinische Befunde stets im klinischen Zusammenhang interpretiert werden müssen, ist jeder Ärztin und jedem Arzt vertraut. Genau dieser Kontextbezug war lange Zeit ganz grundsätzlich eine Schwäche früher KI-Systeme. Erst mit der Einführung der **Transformer-Architektur** im Jahr 2017 [7], die die Grundlage heutiger Large-Language-Mo-

delle wie ChatGPT bildet, wurde eine kontextbezogene Verarbeitung in größerem Maßstab möglich.

Im Unterschied zu klassischen ML-Modellen, die vor allem statische Muster klassifizieren, sind Large-Language-Modelle darauf trainiert, Sequenzen zu verarbeiten, Zusammenhänge im Kontext zu erfassen und das jeweils wahrscheinlich nächste Element vorherzusagen. Sprache – ebenso wie Bildinformationen in multimodalen Systemen – wird hierfür in kleine Einheiten, sogenannte **Token**, zerlegt und in mathematische Vektoren überführt. Semantisch verwandte Begriffe liegen in diesem mehrdimensionalen Vektorraum nahe beieinander; so korrelieren beispielsweise „Herzinfarkt“ und „Myokardinfarkt“ nahezu deckungsgleich.

Das technologische Kernprinzip ist der **Self-Attention-Mechanismus**. Dabei berechnet das Modell für jedes Wort oder Token, welche anderen Elemente des Kontextes für das Verständnis besonders relevant sind. Auf diese Weise kann ein LLM in einem Satz wie „Der Patient erhielt Penicillin, entwickelte danach ein Exanthem und klagte über Juckreiz.“ erkennen, dass sich „danach“ zeitlich auf die Penicillingabe bezieht und, dass Exanthem und Juckreiz als mögliche Folge dieser Behandlung zu interpretieren sind.

Wichtig ist dabei, dass ein LLM nicht im klassischen Sinne auf eine explizite Faktendatenbank zugreift. Es erzeugt seine Antworten **probabilistisch**, also auf Grundlage statistischer Wahrscheinlichkeiten dafür, welches **Token** im jeweiligen Kontext am ehesten folgt. Das erinnert an die intuitive Fähigkeit eines menschlichen Zuhörenden, bei einem Fachvortrag das nächste Wort oder den nächsten Gedankenschritt der Referentin oder des Referenten zu antizipieren.

Die Leistungsfähigkeit speziell für medizinische Fragestellungen trainierter Large-Language-Modelle ist inzwischen bemerkenswert. Wer dies praktisch erproben möchte, kann beispielsweise die werbefinanzierte Plattform **OpenEvidence.com** nutzen. Nachdem man sich – beispielsweise mit einem Foto seines Arztausweises – als einer medizinischen Profession zugehörig verifiziert hat, lassen sich dort klinische Fragen stellen. Obwohl die Benutzeroberfläche englischsprachig ist, unterstützt das System auch deutschsprachige Eingaben und Ausgaben.*

KI-Nutzung durch Patienten

Grundsätzlich ist bei vielen Patientinnen und Patienten eine große Offenheit gegenüber KI-gestützten Anwendungen zu beobachten. Beispielsweise kommen inzwischen auf Large-Language-Modellen basierende Systeme im direkten Patientenkontakt zum Einsatz, etwa im Rahmen der Ersteinschätzung in der Notaufnahme. In einer von uns durchgeführten Studie wurde hierfür ein **multimodales KI-System** verwendet, das über einen Avatar mit Patientinnen und Patienten interagierte. Dabei zeigte sich eine **hohe Patientenzufriedenheit** sowie ein **hohes Maß an Vertrauen** in das System [5].

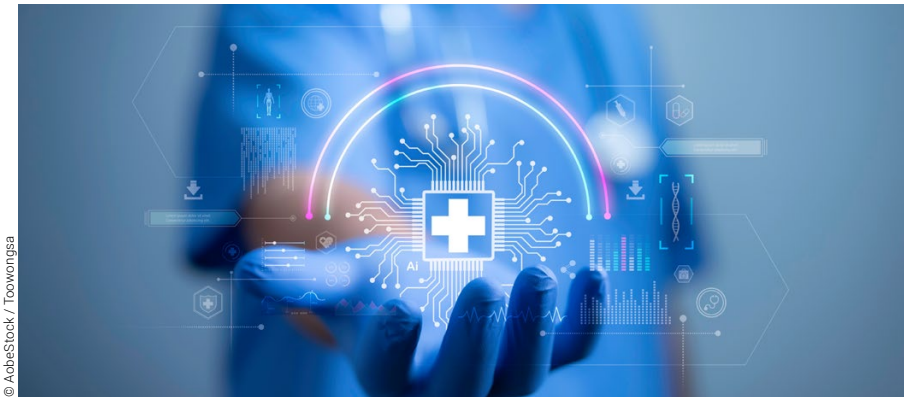
Parallel dazu nutzen Patientinnen und Patienten Large-Language-Modelle zunehmend auch zur **Selbstdiagnose**, ähnlich wie früher Suchmaschinen im Sinne eines „Dr. Google“. Die Ergebnisse sind hier jedoch deutlich problematischer. Eine kürzlich publizierte Studie teilte 1298 Personen in vier Gruppen ein, von denen jeweils drei Zugriff auf ein spezifisches Large-Language-Modell als Chat-Assistenten (GPT-4o, Llama 3, oder Command R+) erhielten, während die vierte Gruppe als **Kontrollgruppe** fungierte und auf die im Alltag üblichen Informationsquellen zurückgreifen sollte, in der Regel Internetsuchmaschinen oder offizielle Gesundheitswebsites [1]. Alle Teilnehmenden bearbeiteten zufällig zugewiesene medizinische Szenarien und mussten entscheiden, welche Erkrankung wahrscheinlich zugrunde lag und welche Behandlungsressource – von Selbstbehandlung bis zum sofortigen Rufen eines Rettungswagens – akut nötig war.

Wurden die Szenarien von den Large-Language-Modellen isoliert, also ohne menschliche Interaktion, bewertet, schnitten die Systeme zunächst recht gut ab: In 94,9 % der Fälle identifizierten sie die relevanten Erkrankungen, und in durchschnittlich 56,3 % der Fälle wählten sie die korrekte Maßnahme. Ganz anders fiel jedoch das Ergebnis aus, wenn Laien diese Systeme tatsächlich nutzten. Die Teilnehmenden, die Large-Language-Modelle verwendeten, schnitten signifikant schlechter ab als die Kontrollgruppe. Sie identifizierten relevante Erkrankungen in < 34,5 % der Fälle. Die Kontrollgruppe, die auf klassische Internetrecherche zurückgriff, hatte eine 1,76-fach höhere Wahrscheinlichkeit, eine relevante Erkrankung korrekt zu erkennen, und war besser darin, kritische Red-Flag-Symptome zu identifizieren [1].

Diese Daten sprechen klar dagegen, dass Patientinnen und Patienten solche Systeme derzeit sicher und zuverlässig für die Selbstdiagnose einsetzen können; die Nutzung kann vielmehr gefährlich sein. Eine plausible Erklärung liegt darin, dass viele Large-Language-Modelle auf medizinischer Fachliteratur trainiert wurden und daher besonders dann leistungsfähig sind, wenn Anfragen in präziser medizinischer Fachsprache formuliert werden können – was bei medizinischen Laien meist nicht gegeben ist.

* Die OpenEvidence-Nutzung ist derzeit von einem EU-gehosteten Server nicht möglich.

KI-Nutzung durch Ärzte



© AdobeStock / Toowongsa

Für Ärztinnen und Ärzte eröffnen sich demgegenüber erhebliche Chancen, insbesondere bei **komplexen** oder **seltenen Krankheitsbildern**, die auch erfahrene Klinikerinnen und Kliniker vor erhebliche diagnostische und therapeutische Herausforderungen stellen können. Keine Ärztin und kein Arzt kann das gesamte medizinische Wissen vollständig überblicken. Gerade hier kann die Zusammenarbeit mit KI, insbesondere mit Large-Language-Modellen, wertvoll sein – ist allerdings nicht ohne Risiken.

Ärztinnen und Ärzte sind in der Lage, die charakteristischen Symptome, Besonderheiten und klinischen Konstellationen eines Krankheitsbildes präzise herauszuarbeiten und in korrekte medizinische Fachsprache zu überführen. Werden Large-Language-Modelle mit derart strukturierten Informationen versorgt, liefern sie häufig sehr leistungsfähige diagnostische und therapeutische Hinweise. In eigenen Tests haben wir sowohl cloudbasierte Systeme wie **ChatGPT von OpenAI**, **Claude von Anthropics** und **Gemini von Google** als auch lokal installierte Modelle wie **Qwen 3.5 von Alibaba** mit den Symptomkonstellationen komplexer seltener Erkrankungen konfrontiert. Dabei konnten selbst sehr seltene Erkrankungen von diesen Systemen mit hoher Trefferquote anhand der vorliegenden Symptome erkannt werden. Dies verdeutlicht das erhebliche Potenzial dieser Technologien.

Für den klinischen Routineeinsatz sind solche Systeme außerhalb von Test- und Forschungsszenarien jedoch derzeit nur eingeschränkt geeignet, solange sie **cloudbasiert in den USA betrieben werden** und damit **nicht DSGVO-konform** eingesetzt werden können. Das Risiko einer Datenschutzverletzung ist dabei keineswegs auf die offensichtliche Weitergabe von Name oder Geburtsdatum beschränkt: Patientinnen und Patienten mit seltenen Erkrankungen oder ungewöhnlichen Symptomkonstellationen lassen sich allein anhand weniger klinischer Merkmale zuverlässig re-identifizieren – eine Pseudonymisierung bietet hier keinen ausreichenden Schutz. Anders stellt sich die Situation bei **lokal installierten Modellen** dar, die keinen Datenverkehr ins Internet erzeugen und patientenbezogene Informationen in einem geschlossenen System verarbeiten. Auch diese Systeme müssen hohe Datenschutzanforderungen erfüllen, ihr Einsatz ist jedoch grundsätzlich datenschutzkonform realisierbar.

In unseren Tests zeigte das frei verfügbare Modell **Qwen 3.5** in der großen **397-Milliarden-Parameter-Variante** eine ausgesprochen gute Leistung bei komplexen medizinischen Fragestellungen. Die Performance lag nur geringfügig unter der moderner Cloudsysteme. Die praktische Nutzung eines solchen Modells erfordert allerdings sehr leistungsfähige Hardware mit **mehr als 400 GB schnellem Videospeicher**, was mit erheblichen Kosten verbunden ist.

Diese Ergebnisse legen nahe, dass KI-Modelle in absehbarer Zeit zu einem ausgesprochen hilfreichen Instrument für Ärztinnen und Ärzte werden können. Es ist zu erwarten, dass kommerzielle Anbieter in den kommenden Monaten und Jahren **datenschutzkonforme** und zugleich **als Medizinprodukt zertifizierte** Lösungen entwickeln werden.

Auch im unmittelbaren **klinischen und praxisbezogenen Alltag** sind KI-Anwendungen inzwischen angekommen. Lungenfunktionsuntersuchungen lassen sich mittlerweile zuverlässig durch KI-Systeme auswerten. Im Bereich der radiologischen Bildanalyse bietet Siemens Healthineers mit dem **AI-Rad Companion** bereits ein System an, das Pathologien im Thorax-CT, beispielsweise fibrotische Veränderungen, automatisiert erkennen kann (vgl. Topic in diesem eJournal, S. 10).

Einen besonders hohen praktischen Nutzen entfalten im Alltag derzeit vor allem Anwendungen, die die **medizinische Dokumentation** vereinfachen und sich in bestehende Praxisverwaltungssysteme integrieren lassen. So kann etwa **Heidi Health** das Arzt-Patienten-Gespräch mithören, transkribieren und daraus nahezu in Echtzeit eine strukturierte Dokumentation erzeugen. Das Unternehmen ist bislang vor allem im angloamerikanischen Raum aktiv. In Europa arbeitet unter anderem **Doctolib** an vergleichbaren Lösungen, die in absehbarer Zeit auch in Deutschland verfügbar sein sollen. Die Berliner Charité-Universitätsmedizin hat

sich Ende 2025 für ein neues Klinik-Informationssystem (KIS) der Firma **Epic Systems** entschieden, das eine KI-gestützte Echtzeit-Sprachanalyse bietet; auch andere KIS-Hersteller bieten solche Funktionalitäten mittlerweile direkt im KIS-System an.

Im Bereich der **Allergologie** befinden sich die meisten KI-Anwendungen hingegen derzeit noch überwiegend im Stadium von Forschungsprojekten. Im Vordergrund stehen hier insbesondere Systeme zur verbesserten **Symptom- und Therapieverlaufskontrolle** sowie Modelle zur **Vorhersage des Erfolgs einer Hyposensibilisierung**. (z. B. [4, 6]).

Ausbildung, Kompetenz und Deskilling

Es ist absehbar, dass sich ärztliches Arbeiten in den kommenden Monaten und Jahren grundlegend verändern wird. Daraus ergibt sich unmittelbar die Notwendigkeit, die **universitäre medizinische Ausbildung** anzupassen und die sichere, kompetente Nutzung von KI-Systemen systematisch in die medizinische Praxislehre zu integrieren.

Dies ist allerdings kein triviales Unterfangen. Schon heute wird im Medizinstudium nicht immer konsequent vermittelt, auf welchen Mechanismen menschliches ärztliches Denken, klinische Entscheidungsfindung und probabilistische Diagnostik im Einzelnen beruhen. Umso anspruchsvoller ist es, zusätzlich die sinnvolle Verzahnung ärztlicher Kognition mit KI-gestützten Entscheidungsprozessen zu lehren.

Darüber hinaus muss definiert werden, **in welchen Bereichen** und **in welcher Form** KI Ärztinnen und Ärzte unterstützen soll, ohne dass dabei die klinische Kernkompetenz verloren geht. Denn der Einsatz von KI-Systemen birgt durchaus die Gefahr eines raschen **Kompetenzverlusts**. Eine Untersuchung zur **Adenomdetektionsrate bei Koloskopien** zeigte, dass in der Endoskopie erfahrene Ärztinnen und Ärzte bereits nach kurzfristigem, regelmäßigem Einsatz von KI-Assistenzsystemen bei Untersuchungen **ohne KI-Unterstützung** signifikant schlechtere Detektionsraten aufwiesen [2]. Dieses Phänomen wird als **Deskilling** bezeichnet.

Deskilling ist kein theoretisches Randproblem, sondern eine hochrelevante Konsequenz zunehmender Automatisierung. Es wird daher notwendig sein, Strategien zu entwickeln, mit denen ärztliche Fähigkeiten durch regelmäßiges Training **ohne Assistenzsysteme** erhalten werden können. Wahrscheinlich wird es hierfür auch **regulatorische Rahmenbedingungen** benötigen.

Fazit

Die Verbindung aus fachärztlicher Erfahrung und der analytischen Breite moderner Sprachmodelle markiert einen potenziell historischen Wendepunkt in der Medizin. Richtig integriert wird KI die Ärztin und den Arzt nicht ersetzen, wohl aber die Diagnostik und Therapiesteuerung erheblich verbessern – weit über den Bereich komplexer oder seltener Erkrankungen hinaus.

Um dieses Potenzial verantwortungsvoll zu nutzen, müssen diese Technologien aktiv mitgestaltet und ihre Anwendung ebenso wie ihre ethischen Implikationen frühzeitig im Medizinstudium und in der ärztlichen Weiterbildung verankert werden. Dabei dürfen die Risiken und Limitationen des KI-Einsatzes nicht übersehen werden. Eine sichere und effektive Nutzung setzt voraus, dass ärztliche Kompetenz auf höchstem Niveau erhalten bleibt und bestimmte Tätigkeiten weiterhin ärztlicher Verantwortung vorbehalten sind, um einem Verlust klinischer Fähigkeiten entgegenzuwirken.

Dr. med. Andreas Jerrentrup

Zentrum für Notfallmedizin
Uniklinikum Marburg
Baldingerstraße
35043 Marburg
jerrentr@staff.uni-marburg.de

Interessenkonflikt:
Andreas Jerrentrup ist als Referent oder Autor tätig für AstraZeneca, Berlin-Chemie, Berufsverband der Deutschen Urologie, CSL Behring, Diabetologen Hessen, Dr. Falk Pharma, Pfizer und Streamed Up.

Literatur

1 Bean AM, Payne RE, Parsons G et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nature Medicine 2026; 32: 609–615

2 Budzyń K, Romańczyk M, Kitala D et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol & Hepatol 2025; 10: 896–903

3 Lång K, Josefsson V, Larsson AM et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol 2023; 24(8): 936–944

4 Mormile I, Granata F, Detoraki A et al. Predictive Response to Immunotherapy Score: A Useful Tool for Identifying Eligible Patients for Allergen Immunotherapy. Biomedicines 2022; 10(5): 971

5 Russ P, Mross PM, Kräling G et al. Feasibility of a multimodal AI-based clinical assessment platform in emergency care: an exploratory pilot study. Front Digit Health 2025; 7: 1657583

6 Sarabu C, Steyaert S, Shah NR. Predicting Environmental Allergies from Real World Data Through a Mobile Study Platform. J Asthma Allergy 2021; 14: 259–264

7 Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Cornell University. Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (v7). arXiv:1706.03762