

## Große Sprachmodelle in der Medizin: Übersichtsarbeit analysiert Risiken und Strategien für sicheren klinischen Einsatz

**Große Sprachmodelle (LLMs), darunter die Modelle hinter ChatGPT und Claude sowie zahlreiche weitere speziell für medizinische Anwendungen entwickelte Systeme, halten Einzug in klinische Arbeitsabläufe. Ihre Nutzung schreitet häufig jedoch schneller voran, als offizielle Sicherheits- und Schutzmaßnahmen Schritt halten können. Ein interdisziplinäres Forschungsteam vom Else Kröner Fresenius Zentrum (EKFZ) für Digitale Gesundheit an der Technischen Universität Dresden (TUD) und dem Dresdner Universitätsklinikum hat gemeinsam mit nationalen und internationalen Kolleg:innen die Risiken beim Einsatz von LLMs in der Medizin systematisch analysiert.**

Die in der Fachzeitschrift Nature veröffentlichte Übersichtsarbeit bündelt Erkenntnisse aus medizinischer KI-Forschung, Cybersicherheit, Regulatorik, Ethik und Verhaltenspsychologie und zeigt Wege für einen vertrauenswürdigen und verantwortungsvollen Einsatz von Künstlicher Intelligenz (KI) in der klinischen Praxis auf.

Große Sprachmodelle können medizinisches Personal in vielen Bereichen unterstützen. Sie werden bereits in der Praxis eingesetzt – häufig jedoch ohne institutionelle Leitlinien oder klare Regeln. Daraus ergeben sich neue Anforderungen für Patientensicherheit, Datenschutz und Verantwortlichkeiten. Die Autor:innen der jetzt veröffentlichten Übersichtsarbeit zeigen, dass Risiken entlang des gesamten Lebenszyklus von KI-Systemen entstehen können: vom Modell-Design über Trainingsdaten und Implementierung bis hin zur Anwendung im klinischen Alltag. Sie unterscheiden dabei drei verschiedene Arten von Risiken:

Sicherheitsrisiken entstehen beispielsweise durch manipulierte Trainingsdaten („Data Poisoning“), gezielte Eingriffe in das Verhalten der Modelle oder sogenannte Prompt Injections. Dabei werden versteckte Anweisungen in Befehle eingebettet, die zu fehlerhaften oder potenziell gefährlichen Ergebnissen führen können – etwa, wenn ein Tumor in einer Gewebeprobe fälschlicherweise nicht erkannt wird. Auch Schwachstellen in der IT-Infrastruktur können sensible Daten von Patient:innen gefährden oder Systeme beeinträchtigen.

Modellinhärente Sicherheitsrisiken ergeben sich aus der Funktionsweise der Systeme: LLMs können plausibel klingende, aber inhaltlich falsche Informationen erzeugen, sogenannte Halluzinationen. Im klinischen Kontext ist dies besonders kritisch, weil falsche Diagnosen oder Empfehlungen die Patientensicherheit gefährden können. Zudem können Modelle ihre Antworten zu stark an die Erwartungen der Nutzenden anpassen und dadurch fehlerhafte Annahmen verstärken.

Risiken in der Mensch-KI-Interaktion betreffen die Anwendung im Alltag: Die Art wie Mediziner:innen die Modelle nutzen, kann klinische Entscheidungen beeinflussen. Überzeugend formulierte Antworten könnten zu einer übermäßigen Abhängigkeit vom System führen („Automation Bias“) oder bestehende Überzeugungen verstärken („Confirmation Bias“). Komplexe oder lange Eingaben könnten zudem die Verlässlichkeit der Ergebnisse verringern.

Der Artikel weist darüber hinaus auf strukturelle und ethische Herausforderungen hin. Viele

Systeme werden nicht lokal betrieben, was relevante Fragen zu Datenkontrolle und Datenschutz aufwirft. Zudem findet vielerorts bereits eine informelle Nutzung von LLMs im klinischen Alltag statt. Man spricht von einer Schattennutzung, oft ohne entsprechende Sicherheitsmaßnahmen.

„Unsere Analyse zeigt, dass große Sprachmodelle klinische Arbeitsabläufe sinnvoll unterstützen können. Ihr sicherer Einsatz ist jedoch nicht selbstverständlich. Risiken entstehen an vielen Stellen und müssen vor sowie parallel zur klinischen Umsetzung systematisch und umfassend adressiert werden.“, sagt Dr. Jan Clusmann, Postdoktorand in der Forschungsgruppe von Prof. Jakob N. Kather am EKfZ für Digitale Gesundheit an der TUD und Erstautor der Publikation.

### **Strategien für einen sicheren Einsatz**

Um Risiken zu reduzieren, schlagen die Autor:innen verschiedene Maßnahmen vor. Dazu zählen sichere Entwicklungsprozesse, sorgfältige Aufbereitung von Trainingsdaten, systematische Evaluation und kontinuierliches Monitoring von Modellen. Ebenso notwendig sind klar definierte Zuständigkeiten innerhalb von Gesundheitseinrichtungen. Sicherheit ist dabei nicht allein eine technische Frage, sondern erfordert ein abgestimmtes Vorgehen zwischen Forschung, klinischer Praxis und Regulierung. Die menschliche Kontrolle bleibt zentral, so die Autor:innen.

„Die Entwicklung medizinischer KI endet nicht mit leistungsfähigen Modellen. Ebenso wichtig ist es, Sicherheit, Transparenz und Nutzen für die klinische Praxis wissenschaftlich zu untersuchen. Dass unsere Forschenden hierfür evidenzbasierte Grundlagen schaffen, ist ein wichtiger Beitrag für eine verantwortungsvolle Digitalisierung der Medizin“, sagt Prof. Esther Troost, Dekanin der Medizinischen Fakultät Carl Gustav Carus der TU Dresden.

Konkret empfehlen die Forschenden den Aufbau klarer Strukturen, etwa spezialisierter Teams vor Ort, die den Einsatz von KI-Systemen im Klinikalltag koordinieren und überwachen. Solche zentralen Einheiten – sogenannte Security Operations Centers (SOCs) for AI – ermöglichen es, Vorfälle institutionsübergreifend zu erkennen und koordinierte Gegenmaßnahmen zu ergreifen.

„KI wird schon jetzt im Gesundheitswesen eingesetzt – häufig ohne formale Aufsicht. Entscheidend ist daher, wie wir diese Systeme transparent, robust und im Einklang mit klinischer Verantwortung implementieren“, sagt Prof. Jakob N. Kather, Professor für Klinische Künstliche Intelligenz am EKfZ für Digitale Gesundheit an der TUD, Arzt am Universitätsklinikum Dresden sowie Wissenschaftler am Nationalen Centrum für Tumorforschung (NCT) Heidelberg.

### **Regulierung an dynamische KI-Systeme anpassen**

Die Analyse zeigt auch Lücken in der aktuellen Regulierung auf. Nur ein kleiner Teil der KI-Systeme ist bislang als Medizinprodukt zugelassen. Der bestehende regulatorische Rahmen ist nicht auf anpassungsfähige Technologien ausgelegt, die sich kontinuierlich weiterentwickeln.

„Um die Patientensicherheit zu gewährleisten und zugleich einen zeitnahen Nutzen für Patient:innen sowie das Gesundheitssystem durch KI-Systeme zu ermöglichen, braucht es geeignete regulatorische Ansätze, die eine konsistente Bewertung und kontinuierliche Überwachung sicherstellen. Technologische Entwicklung sowie Aufsichts- und Überwachungsmechanismen müssen enger verzahnt werden, um das Potenzial großer Sprachmodelle sicher zu nutzen“, sagt Prof. Stephen Gilbert, Professor für Medical Device Regulatory Science am EKfZ für Digitale Gesundheit an der TU Dresden.

### **Die Publikation entstand in Zusammenarbeit an folgenden nationalen und internationalen Institutionen:**

Else Kröner Fresenius Zentrum (EKFZ) für Digitale Gesundheit, Medizinische Fakultät und Universitätsklinikum Carl Gustav Carus, Technische Universität Dresden, Universitätsklinikum RWTH Aachen, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg, Nationales Centrum für Tumorerkrankungen (NCT) Heidelberg, Universitätsklinikum Heidelberg, Medizinische Fakultät der Universität Heidelberg, Medizinische Fakultät Mannheim, Universitätsmedizin Mainz, Purdue University West Lafayette, University of Pennsylvania sowie das Patientenrechte-Netzwerk The Light Collective Eugene.

## **Publikation**

Clusmann J, Freyer O, Ostermann M, Ferber D, Ghaffari Laleh N, Hilgers L, Kolbinger FR, Schneider CV, Downing A, Wekenborg MK, Gilbert S, Foersch S, Truhn D, Wiest IC, Kather JN. Safety and security of large language models in healthcare. Nature, 2026.

Link: <https://www.nature.com/articles/s41586-026-10687-1>

## **Else Kröner Fresenius Zentrum (EKFZ) für Digitale Gesundheit**

Das EKFZ für Digitale Gesundheit an der Medizinischen Fakultät der Technischen Universität Dresden (TUD) und dem Universitätsklinikum Carl Gustav Carus Dresden wurde im September 2019 gegründet. Es wird mit einer Fördersumme von 40 Millionen Euro für eine Laufzeit von zehn Jahren von der Else Kröner-Fresenius-Stiftung gefördert. Das Zentrum konzentriert seine Forschungsaktivitäten auf innovative, medizinische und digitale Technologien an der direkten Schnittstelle zu den Patientinnen und Patienten. Das Ziel ist dabei, das Potenzial der Digitalisierung in der Medizin voll auszuschöpfen, um die Gesundheitsversorgung, die medizinische Forschung und die klinische Praxis nachhaltig zu verbessern.