Empathie, Nutzen und Korrektheit von Large-Language-Models wie ChatGPT in der medizinischen Entscheidungsfindung

Ein Experiment mit Einbindung von Ärztinnen und Ärzten sowie Patientinnen und Patienten

Problem

- Sog. „Large Language Models“ (LLMs), basierend auf maschinellem Lernen und künstlicher Intelligenz, haben Einzug in das tägliche Leben gehalten (s.u.a. Siri, Alexa)
- ChatGPT (https://chatgpt.com) ist die derzeit wohl bekannteste KI-basierte Plattform für Wissensgenerierung, -zusammenführung und -umsetzung
- Es bestehen erhebliche Unsicherheiten und Ängste sowohl bei Anbietern von Gesundheitsdienstleistungen als auch Patientinnen und Patienten, ob LLMs / KI zukünftig die menschliche (ärztliche) Beratung und Kompetenz ersetzen können oder werden.

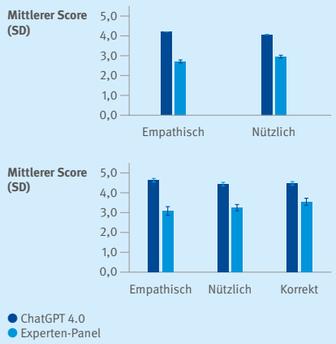
Versuchsaufbau

- Experiment mit 100 öffentlich verfügbaren Fragen von Patientinnen und Patienten und Antworten von Ärztinnen und Ärzte in Unfallchirurgie, Allgemeinchirurgie, Pädiatrie, innerer Medizin und HNO
- Erneute Beantwortung der Fragen durch ChatGPT 4.0
- 64 Patientinnen und Patienten (29 Männer, 35 Frauen, mittleres Alter 45 [Spanne 16–76] Jahre), stationär und ambulant in der BG Klinik Ludwigshafen
- 5×3 Ärztinnen und Ärzte der o.g. Disziplinen mit einer mittleren Erfahrungsdauer von 11 (Spanne 5–34) Jahren
- Auf einer Skala von 1 (gering) bis 5 (hoch) wurden Empathie, Nutzen und Korrektheit der Aussagen eingeschätzt

Ergebnis

Patientinnen und Patienten bewerteten Chatbot-Antworten als empathischer und nützlicher als diejenigen von Menschen

Ärztinnen und Ärzte bestätigten diesen Trend



www.bg-kliniken.de

BG Klinik Ludwigshafen

Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H.

Journal of Medical Internet Research

"Doctor ChatGPT, Can You Help Me?" The Patient's Perspective: Cross-Sectional Study. *J Med Internet Res.* 2024 Oct 1;26:e58831. doi: 10.2196/58831. PMID: 39352738; PMCID: PMC11480680.

Was bisher bekannt ist

Large Language Modelle (LLM) haben Einzug in zahlreiche Lebensbereiche gehalten. Sie stellen eine Form der generativen künstlichen Intelligenz (KI) für textbasierte Inhalte dar, „verstehen“ humane Fragen und beantworten diese unter Nutzung von Milliarden von Informationseinheiten aus allen elektronischen Quellen und einer neuronalen Netzwerkarchitektur.

LLM verknüpfen dabei wahrscheinlichste Wortgefüge und deren Kombinationen. Sie ermöglichen sowohl Vertreterinnen und Vertretern der Gesundheitsberufe als auch Patientinnen und Patienten, Antworten zu einem bestimmten Thema erhalten. Ob diese immer fachlich korrekt sind, den aktuellen Stand des medizinischen Wissens abbilden, oder diese in einen sinnvollen Kontext bringen, ist derzeit noch unklar. Ein typisches Phänomen ist das sogenannte „Halluzinieren“ – die Antworten eines Chatbots sind dabei zwar in sich schlüssig, beantworten aber die von Menschen gestellten Fragen nicht. LLM, derzeit vorrangig repräsentiert durch ChatGPT (Open AI Inc., San Francisco, USA), werden stetig aktualisiert und durch neue Versionen in kurzen Abständen erweitert. In verschiedenen Gebieten der Medizin herrscht Unklarheit, wie medizinische Laien und Betroffene Antworten einer KI auf ihre Anfragen zu spezifischen Gesundheitsproblemen und Behandlungsempfehlungen bewerten und für eine Entscheidungsfindung nutzen. Auch die medizinische Korrektheit der KI-basierten Antworten wurde bisher nur unzureichend untersucht.

Studiendesign und Resultate

Für eine experimentelle Untersuchung wurden 100 reale Fragen von Patientinnen und Patienten aus einem Onlineforum samt entsprechender Antworten eines ärztlichen Experten-Panels aus den Bereichen Traumatologie, Allgemeinchirurgie, HNO, Pädiatrie und Innerer Medizin ausgewählt. Die Untersucher ließen die Fragen für dieses Experiment erneut von ChatGPT 4.0 beantworten und kreierten Pakete aus 10 × 10 Fragen und Antworten. Diese wurden insgesamt 64 Patientinnen und Patienten (mittleres Alter 46 ±16 Jahre, 29 Männer, 35 Frauen) und 15 Fachärztinnen und Fachärzten mit einer mittleren Berufserfahrung von 11 (Spanne, 5–34) Jahren vorgelegt. Ohne Kenntnis, ob die Antworten vom Chatbot oder einem Menschen stammten, sollten diese auf einer Skala von 1 (gering) bis 5 (hoch) Empathie und Nutzen (im Sinne der Entscheidung zugunsten einer bestimmten Intervention) bewerten. Ärztinnen und Ärzte beurteilten zudem die medizinische Korrektheit der Aussagen und mögliche schädliche Therapieempfehlungen.

Patientinnen und Patienten schätzten Empathie und Nutzen der Chatbot-Antworten deutlich höher als diejenigen der menschlichen ein (4,2 versus 2,7 und 4,1 versus 3,0). Dies wurde von Ärztinnen und Ärzten ebenso gesehen (4,7 versus 3,1 und 4,5 versus 3,3). Auch die medizinische Korrektheit wurde besser beurteilt (4,5 versus 3,6).

Alarmierend war, dass Antworten des Chatbots, welche von Ärztinnen und Ärzten als potenziell gefährlich eingestuft wurden, von Patientinnen und Patienten als ebenso nützlich und empathisch eingeschätzt wurden wie die als ungefährlich eingestuften Antworten (4,2 versus 4,2 und 4,0 versus 4,1). ChatGPT 4.0 führte theoretisch zu vier Überdiagnosen bzw. zur Empfehlung unnötiger Interventionen, acht verzögerten Diagnosen oder verspäteten Behandlungen und drei unzureichenden Aufklärungen.

Bedeutung für die klinische Versorgung und Forschung in den BG Kliniken

ChatGPT und andere Algorithmen sind im Stande, unter Nutzung einer humanen Dateneingabe bestimmte Voraussagen für den Eintritt von Ereignissen zu treffen und mit Ratsuchenden sinnvoll zu kommunizieren. Sie stellen vielversprechende Werkzeuge zur Unterstützung der medizinischen Entscheidungsfindung dar. Ohne menschliche Aufsicht über potenzielle Risiken, falsche, unvollständige oder verzerrte Informationen oder gar kriminelle Intention dürfen sie jedoch nicht ungefiltert genutzt werden.

https://www.jmir.org/2024/1/e58831