

Dr. KI unter der Lupe

Neue Studien der TU Berlin zeigen Stärken und Schwächen von KI-gestützten Tools zur digitalen Gesundheitsberatung wie ChatGPT oder Symptom-Checker-Apps

Patient*innen greifen immer häufiger auf digitale Tools zurück, um Krankheiten zu erkennen und Handlungsempfehlungen zu erhalten. Zwei aktuelle Studien der TU Berlin haben nun die Qualität und Wirksamkeit solcher digitalen Gesundheitsempfehlungen untersucht. Die Ergebnisse zeigen sowohl Potenziale als auch Risiken auf. Die Studien sind in den Fachzeitschriften Journal Scientific Reports und npj Health Systems Journal erschienen.

In der ersten Studie <https://www.nature.com/articles/s41598-024-83844-z> entwickelte ein Team um Dr. Marvin Kopka aus dem Fachgebiet Arbeitswissenschaft an der TU Berlin eine neue Testmethode, um die Genauigkeit von Gesundheitsempfehlungen durch ChatGPT und andere KI-Modelle wie Meta's LLaMa sowie speziell entwickelte Symptom-Checker-Apps, die Symptome abfragen und darauf basierend Handlungsempfehlungen geben, zu bewerten. Während frühere Tests auf idealisierten Lehrbuchfällen basierten, die in der Realität so kaum vorkommen, setzt die neue Methode auf echte Patientenfälle, mit denen die Wissenschaftler*innen der TU Berlin verschiedene digitale Tools testeten. Dadurch lässt sich realistischer einschätzen, wie präzise und hilfreich die digitalen Tools in der Praxis sind. „Man kann unsere standardisierte Methode als eine Art ‚Stiftung Warentest‘ ansehen, da wir mit ihr die Genauigkeit von verschiedenen Apps vergleichen können, aber auch ihre Stärken und Schwächen finden“, sagt Studienleiter Marvin Kopka.

Symptom-Checker-Apps deutlich hilfreicher

Die Ergebnisse der neu entwickelten Bewertungsmethode zeigen, dass Symptom-Checker-Apps für Laien deutlich hilfreicher sind als ChatGPT, insbesondere wenn es um die Unterscheidung zwischen harmlosen und ernstesten Symptomen geht. Während ChatGPT fast jeden Fall als Notfall oder hochdringlich einstuft, können die spezialisierten Apps in den meisten Fällen fundierte und angemessenere Empfehlungen geben. Was auch schon andere Studien belegen: ChatGPT kann Krankheiten gut diagnostizieren, wenn Laborwerte oder Untersuchungsergebnisse vorliegen. Da diese zu Hause meist fehlen, bleibt die Diagnose aber oft ungenau und die vom Modell vorgeschlagene Liste mit mehreren möglichen Erkrankungen hilft Laien wenig. Entscheidender sind Handlungsempfehlungen wie „Geh zum Arzt“ oder „Ruf die 112“ – hier schneidet ChatGPT jedoch schlecht ab, da es fast jeden Fall als behandlungsbedürftig einstuft, selbst bei harmlosen Symptomen.

Laien erkennen medizinische Notfälle meist zuverlässig

Was die Forscher*innen auch herausfanden: Laien erkennen medizinische Notfälle meist zuverlässig und rufen in ernstesten Fällen, wie bei einer schweren Kopfverletzung mit Erbrechen und Schwindel, den Rettungsdienst. Schwerer fällt es ihnen jedoch, harmlose Symptome richtig einzuschätzen. So neigen viele dazu, bei leichten Beschwerden wie kurzfristigem Durchfall oder einer kleinen Hautveränderung vorschnell ärztliche Hilfe in Anspruch zu nehmen, obwohl dies oft nicht nötig wäre. „Dass immer mehr Menschen ChatGPT für medizinische Ratschläge nutzen, ist schädlich für das Gesundheitssystem. Die KI motiviert die Nutzer*innen häufig dazu, bei den kleinsten Symptomen sofort einen Arzt oder die Notaufnahme aufzusuchen. Das kann zu einer massiven

Überlastung führen“, warnt Studienleiter Dr. Marvin Kopka.

Nutzer*innen übernehmen digitale Empfehlungen nicht unkritisch

Die zweite Studie <https://www.nature.com/articles/s44401-024-00008-x> verglich nicht nur Menschen und Technik, sondern untersuchte, wie genau Menschen die Empfehlungen von ChatGPT und Symptom-Checker-Apps in ihre eigenen Entscheidungen einbeziehen. Dabei zeigte sich, dass Nutzer*innen die Empfehlungen nicht unkritisch übernehmen, sondern mit anderen Quellen wie Google-Suchen, dem Rat von Freunden oder weiteren Apps abgleichen. „Auf der anderen Seite gibt es auch Fälle, in denen Patient*innen zu viele und teils unverständliche Informationen von den digitalen Tools bekommen, die sie nicht einordnen können. Das erzeugt Angst und sie suchen dann den Expertenrat in der Notaufnahme oder beim Hausarzt – auch bei harmlosen Beschwerden, so wie es ihnen ChatGPT empfiehlt“, so Kopka.

Vor einer quantitativen Untersuchung mit 600 Proband*innen wurden in der zweiten Studie zunächst 24 Personen bei der Nutzung beobachtet und anschließend ein Modell aufgestellt, wie sie Entscheidungen mit Hilfe von ChatGPT und Apps treffen. Die Auswertung ergab erneut, dass ChatGPT die Selbstversorgung erschwert und die Zahl unnötiger Arztbesuche erhöht. Demgegenüber konnten gut funktionierende Symptom-Checker-Apps dazu beitragen, dass Nutzer*innen sich in passenden Fällen für die Selbstversorgung entschieden und so zur Entlastung des Gesundheitssystems beitrugen. „ChatGPT hat viele sinnvolle Anwendungsfälle, aber zur Entscheidung, ob ich zum Arzt gehen sollte, ist es nicht geeignet – dafür ist es viel zu ungenau“, resümiert Kopka. „Wir sollten uns eher fragen, ob eine App uns hilft, gute Entscheidungen zu treffen, statt von ihrer Perfektion zu erwarten. Denn Menschen treffen bereits jetzt in den meisten Fällen sichere und vernünftige Entscheidungen. In einigen Situationen können Sie jedoch von den Apps profitieren.“

Die Studien verdeutlichen, dass digitale Tools sinnvoll eingesetzt werden können, um Patient*innen in ihrer Entscheidungsfindung zu unterstützen. Speziell entwickelte Symptom-Checker-Apps erweisen sich dabei aktuell als hilfreicher als generative KI-Modelle wie ChatGPT. Dennoch bleibt laut der Wissenschaftler*innen ein kritischer Umgang mit digitalen Empfehlungen entscheidend, um Fehleinschätzungen und eine unnötige Belastung des Gesundheitssystems zu vermeiden.

Publikationen:

Journal Scientific Reports: The RepVig framework for designing use-case specific representative vignettes and evaluating triage accuracy of laypeople and symptom assessment applications <https://www.nature.com/articles/s41598-024-83844-z>

npj Health Systems Journal: Technology-Supported Self-Triage Decision Making <https://www.nature.com/articles/s44401-024-00008-x>