Einsichten eines Wissenschaftsnarren (73)

KI in der Medizin: Hybris, Hype, Halbwissenschaft

KI hat das Potenzial, die Medizin zu revolutionieren. Jedoch stehen zumindest drei Dinge einer wirklich evidenzbasierten KI im Weg: eine unbedachte „Move fast, break things"-Mentalität, dazu industriefreundliches Lobbying – sowie nicht zuletzt wissenschaftliche Defizite in Transparenz, Validierung und Bias-Reduktion.

„Chatbot schlägt Ärzte bei der Diagnose!" So schallte es aus den Medien von *X* bis *New York Times*. Eine randomisiert kontrollierte Studie einer illustren Schar von US-Autoren aus Stanford, Harvard und anderen Edelschmieden hatte angeblich gezeigt, dass ChatGPT korrektere Diagnosen stellt als Mediziner in den genannten weltbekannten Unikliniken. Nicht nur sei die künstliche Intelligenz (KI) mit 92 Prozent deutlich akkurater als die Ärzte mit 74 Prozent gewesen – nein, selbst wenn Letztere den Chatbot nutzen durften, seien sie mit 76 Prozent nur marginal besser geworden. Das Fazit ist also nicht nur, dass KI bessere Diagnosen stellt als Ärzte. Vielmehr können diese zudem nicht mit Chatbots umgehen – und lassen sich von der KI auch nicht belehren, obwohl sie es besser kann!

»Dies ist bereits das vierte Mal, dass die KI-Sau durchs Dorf getrieben wird.«

Dies ist ein wunderbares Beispiel. Allerdings nicht für die enormen Fähigkeiten, die KI in der Medizin bereits vermeintlich hat, sondern für den extremen Hype rund um KI wie auch für die mangelnde Qualität entsprechender Studien. Und deshalb muss sich der Narr dem Thema nochmals annehmen, nachdem er sich vor einiger Zeit schon mal ganz prinzipiell zur angeblichen „Intelligenz" von KI geäußert hat (*LJ* 6/23: 18-21).

Zunächst zum Hype. Der ist schnell abgehandelt, denn es dürfte mittlerweile auch den naivsten Zeitgenossen aufgefallen sein, wie überspannt die Versprechungen zu KI im Gesundheitssystem sind. Nur zwei Beispiele, *pars pro toto*: Deloitte, eine Consulting-Firma, prognostiziert die Rettung von 400.000 Leben, die Einsparung von 200 Milliarden Euro und 1,8 Milliarden Arbeitsstunden allein in der EU – und das natürlich pro Jahr (Zitate und weiterführende Literatur wie immer unter *dirnagl.com/lj*). Und im Januar konfabulierte OpenAI-Chef Sam Altman bei der Vorstellung des Stargate-Projekts durch Donald Trump („500 Milliarden Dollar für KI in USA"): „KI wird helfen, Krankheiten in noch nie dagewesener Geschwindigkeit zu heilen. Wir werden staunen, wie schnell wir diesen und jenen Krebs sowie Herzkrankheiten heilen. Ganz generell denke ich, dass diese Fähigkeit, […] Krankheiten in rasantem Tempo zu heilen, eine der wichtigsten Errungenschaften dieser Technologie sein wird."

Die Älteren unter Ihnen werden sich vielleicht erinnern, dass dies bereits das vierte Mal ist, dass die KI-Sau durchs Dorf getrieben wird: Marvin Minsky, ein KI-Pionier war sich 1967 sicher, dass „innerhalb einer Generation das Problem der KI gelöst sein wird". In den Achtzigerjahren des vorigen Jahrhunderts glaubte man, dass „Expertensysteme bald Ärzte und Juristen ersetzen werden". Als das Machine-Learning (ML)-Tool „Deep Blue" dann den damaligen Weltmeister im Schach Garri Kasparov schlug, fühlte man sich der allgemeinen artifiziellen Intelligenz (AGI) von neuronalen Netzen ganz nah. Folglich hat ChatGPT natürlich einen Punkt, wenn es die Geschichte der KI selbstbewusst folgendermaßen sieht: „KI-Hype kommt und geht, aber der Fortschritt bleibt."

Die Wissenschaft kriegt gleich ihr Fett weg, zuvor aber: Wie viel KI ist denn eigentlich schon „auf der Straße", wie viel KI nutzen wir heute in der Medizin? Schließlich versucht man ja schon seit deutlich mehr als einem Jahrzehnt sehr intensiv und mit noch mehr Geld KI in medizinische Anwendungen zu bringen. Man denke nur an „Watson for Oncology" von IBM: Gestartet 2010, begraben mitsamt der ganzen KI-Health-Sparte von IBM im Jahr 2022.

Derzeit gibt es rund tausend von der US-amerikanischen Food and Drug Administration (FDA) zugelassene KI- und ML-Algorithmen – diese hauptsächlich in der medizinischen Bildgebung. Klingt nach viel, ist es aber nicht, wenn man bedenkt, wie viele Ressourcen hineingesteckt wurden und wie lange man schon dran ist. Außerdem findet sich die Mehrheit der FDA-Zulassungen nur in wenigen Fachgebieten wie Radiologie oder Kardiologie, mit stark überlappenden Funktionalitäten.

»Empirische Belege für die Verbesserung von Diagnostik und Therapie durch KI sind rar.«

Viel interessanter ist jedoch die Frage, wie viele KI-Tools es bereits mit hohem Evidenz-Level in die Guidelines von medizinischen Fachgesellschaften geschafft haben. Diese lassen sich leider an nur zwei Händen abzählen. Zudem werden sie nicht in den Guidelines empfohlen, weil sie die Patientengesundheit verbessern – sondern weil sie einige ärztliche Tätigkeiten effektiver machen, also schlichtweg Zeit sparen. Daran ist nichts verkehrt, aber richtig prickelnd ist es noch nicht.

Dass so wenig KI in den Guidelines empfohlen wird, ist jedoch nicht verwunderlich. Denn empirische Belege für die Kosteneffizienz, geschweige denn für die Verbesserung von Diagnostik und Therapie durch KI im Gesundheitswesen sind rar – und wo vorhanden, häufig methodisch unzureichend. Die meisten Studien konzentrieren sich auf technische Leistungskennzahlen oder klinische Machbarkeit, sind also im Wesentlichen Proof of Concept (PoC). Robuste gesundheitsökonomische Bewertungen fehlen.

Beispielsweise fand ein systematischer Review nur 86 randomisiert kontrollierte Studien im Feld, von denen allerdings über zwei Drittel zu klein oder methodisch fraglich wa-

Sämtliche Folgen der „Einsichten eines Wissenschaftsnarren" gibt es unter
www.laborjournal.de/rubric/narr

ren. Ohne solche Studien in der erforderlichen Qualität wissen wir jedoch nicht, ob einzelne KI-Tools tatsächlich Kosteneinsparungen bringen oder ärztliche Entscheidungen und klinische Endresultate verbessern können. Oder vielmehr vielleicht sogar verschlechtern.

Und damit sind wir bei der KI-Wissenschaft. Diese steckt, wer hätte das gedacht, wie so manches andere Forschungsfeld in einer Reproduzierbarkeitskrise. Nur rund fünf Prozent der KI-Forscher teilen ihre Quellcodes, und weniger als ein Drittel stellen ihre Trainings- oder Validierungsdaten zur Verfügung. So kann man gar nicht versuchen, etwas zu reproduzieren. Und dort, wo es möglich war und versucht wurde, waren weniger als ein Drittel der Schlüsselresultate von KI-Studien reproduzierbar. Wie aber soll etwas, das man nicht wiederholen kann, zur soliden Grundlage für die Entwicklung nützlicher Tools in der Biomedizin werden?

Reproduzierbarkeit ist ein Problem quer durch die Wissenschaft. Aber die Reproduzierbarkeit von KI-generierten Ergebnissen steht zusätzlich vor zahlreichen für diese Techniken spezifischen Herausforderungen. Zufälligkeit und Stochastizität können dazu führen, dass Algorithmen im Deep Learning unterschiedliche Ergebnisse liefern. Ein Mangel an Standardisierung in der Vorverarbeitung, beispielsweise in der Datenkennzeichnung für die Klassifizierung, kann die Modellleistung erheblich beeinflussen. Nicht-deterministische Hardware- und Softwarebedingungen, wie etwa Unterschiede zwischen den Prozessoren verschiedener Hersteller, können ebenfalls zu abweichenden Resultaten führen.

Hinzu kommt, dass Versionsprobleme, etwa die Umstellung von verschiedenen Versionen einer ML-Bibliothek, signifikante Unterschiede in den Ergebnissen verursachen können. Auch die Verfügbarkeit und Variabilität von Datensätzen stellt ein Problem dar, da proprietäre Gesundheitsdatensätze oft nicht zugänglich sind, wodurch unabhängige Replikationen verhindert werden. Ein weiteres Problem besteht im Überanpassen an spezifische Trainingsdatensätze. Die Interpretation der Ergebnisse wird gerade durch eine übermäßige Abhängigkeit von denselben wenigen Datensätzen erschwert. Schließlich entstehen Verzerrungen durch selektive Berichterstattung, wenn nur die besten Versuchsergebnisse veröffentlicht werden, während weniger erfolgreiche Durchläufe unerwähnt bleiben.

Weiterhin sind KI-Algorithmen in der Regel Black Boxes, deren Ergebnisse oft nicht nachvollziehbar sind. Maschinelle Lernmethoden sind atheoretisch, assoziativ und häufig undurchsichtig. Dadurch wird die Erklärbarkeit („Explainability") zu einer zentralen Herausforderung für KI. Wenn Nutzer die Entscheidungswege nicht verstehen, sind Fehler und Verzerrungen schwerer zu erkennen und zu korrigieren.

Menschen fällt es generell schwerer, Vorhersagen oder Empfehlungen zu akzeptieren, wenn sie nicht nachvollziehbar sind. Eine besondere Herausforderung ergibt sich daher aus der Wechselwirkung zwischen zwei gegensätzlichen psychologischen Phänomenen: der „Algorithmic Aversion", also der Skepsis gegenüber algorithmischen Entscheidungen, und dem „Automation Bias", der dazu führt, dass Menschen automatisierten Systemen oft blind vertrauen. Während einige Nutzer KI-gestützte Entscheidungen kritisch hinterfragen oder ablehnen, neigen andere dazu, sie ungeprüft zu akzeptieren und sich weniger auf ihr eigenes Urteilsvermögen oder eine manuelle Überprüfung zu verlassen. Diese Dynamik macht den verantwortungsvollen Einsatz von KI umso komplexer – dies noch mehr, als ihre Ergebnisse nicht nachvollziehbar sind.

Dazu kommt, dass ein substanzieller Korpus des „kausalen Wissens" der Medizin sich im Nachhinein als falsch herausstellt: Die mit offensichtlich falschen Theorien begründeten, aber empirisch mit randomisiert kontrollierten Studien als erfolgreich belegten Therapien werden trotzdem weiter eingesetzt. Und meist finden wir gleich eine neue, passendere Theorie, wodurch die Erklärbarkeit scheinbar wieder hergestellt ist. Eine generelle Forderung nach vollständiger Erklärbarkeit von KI-Entscheidungen in der Medizin ist daher vermutlich unbegründet, könnte gar schädlich sein. In jedem Fall erfordert jedes klinische KI-Tool eine individuelle Bewertung der Erklärbarkeitsanforderungen, was die Sache nicht einfacher macht.

»KI-Modelle bergen das Risiko, bestehende Verzerrungen in der Forschung zu verstärken.«

Zu den Kernproblemen der ML-basierten KI zählt weiterhin auch der Daten-Drift. Er tritt auf, wenn sich die Daten mit der Welt um sie herum weiterentwickeln, der Algorithmus jedoch in dem Zeitraum verbleibt, in dem er trainiert wurde. In der Medizin passiert das ständig. Die medizinische Praxis ändert sich, dazu die Bevölkerungsstruktur – und noch vieles mehr. Da wird die KI sogar potenziell zum Opfer ihres eigenen Erfolges: Sollte sie dazu beigetragen haben, Diagnosen oder Therapien erfolgreich zu verbessern, ist die Wahrscheinlichkeit hoch, dass ihre Vorhersagen und Empfehlungen dadurch schlechter werden.

Wenn das passiert, kann das viele Menschenleben kosten – wie geschehen bei Epic's Sepsis Prediction Model. Dieses wurde bei einem der riesigen Klinikkonzerne in den USA eingesetzt, um das Risiko für die Entwicklung einer Sepsis vorherzusagen und entsprechend zu therapieren. Das hat eine Weile gut funktioniert, bis sich die Kodierung von Sepsis im Klassifikationssystem für medizinische Diagnosen (ICD) veränderte und der Konzern zusätzliche Krankenhäuser gekauft hatte, die nicht im Trainingsdatensatz waren.

Dies ist ebenso ein Beispiel für die Generalisierungsprobleme von KI-gestützten Gesundheitssystemen. Auch IBM Watson for Oncology funktionierte eine Weile ganz gut am Memorial Sloan Kettering Cancer Center, konnte aber nicht auf andere Krankenhäuser übertragen werden. Optums Healthcare KI hingegen diskriminierte schwarze Patienten bei der Risikobewertung. Googles Retinopathy Detection Model funktionierte in Studien, scheiterte anschließend jedoch im Praxiseinsatz in Thailand. Die NHS- und Babylon-Health-KI gab irreführende oder unsichere medizinische Ratschläge. COVID-19-KI-Modelle versagten unter realen Bedingungen. Googles Brustkrebs- und Stanfords Pneumonie-Modelle erzielten gute Ergebnisse in Tests, aber nicht in klinischen Einsätzen. Google Flu Trends scheiterte spektakulär bei der Grippevorhersage. Eine KI zur Hautkrebsdiagnose schnitt bei dunk-

Foto: BIH/Thomas Rafalzyk

Ulrich Dirnagl

ist experimenteller Neurologe an der Berliner Charité und Gründungsdirektor des QUEST Center for Responsible Research am Berlin Institute of Health. Für seine Kolumne schlüpft er in die Rolle eines „Wissenschaftsnarren" – um mit Lust und Laune dem Forschungsbetrieb so manche Nase zu drehen.

ler Haut schlechter ab. PathAIs Diagnostik-KI führte zu unterschiedlichen und fehlerhaften Diagnosen von Krebs. Die Liste ließe sich beliebig fortsetzen.

In alledem steckt das Bias-Problem: Sämtliche KI-Modelle, insbesondere aber die momentan so populären großen Sprachmodelle (LLMs), bergen das Risiko, bestehende Verzerrungen in der Forschung zu verstärken. Dies passiert auf allen Ebenen des KI-Lebenszyklus: vom Datensammeln und der Annotation über die eigentliche Modellentwicklung bis hin zum „Deployment" und der Evaluierung. Viele veröffentlichte wissenschaftliche Informationen sind falsch, veraltet oder voreingenommen. Da KI-Modelle auf diesen teils fehlerhaften Daten trainiert werden, verbreiten sie deren Mängel weiter und verstärken sie sogar. Eine zentrale Herausforderung besteht darin, zwischen vertrauenswürdigen und weniger glaubwürdigen Informationsquellen sowie zwischen voreingenommenen und neutralen Studiendesigns zu unterscheiden. Das gelingt selbst Experten häufig nicht.

»Über 80 Prozent der KI-Richtlinien zeigen auf, was *gemacht werden soll – aber nicht* wie.*«*

Aber gibt es nicht bereits eine Lösung für all diese Probleme? Man müsste KI doch nur „trustworthy", also vertrauenswürdig machen. Und hat nicht bereits 2019 eine High-Level Expert Group on Artificial Intelligence der Europäischen Kommission „Ethikrichtlinien für vertrauenswürdige KI" erstellt? Baut nicht der 2024 verabschiedete EU Artificial Intelligence Act (144 Seiten!) darauf auf? Ist Trustworthy KI damit nicht sogar gesetzlich kodifiziert, zumindest in der EU? Es gibt doch auch aktuelle Stellungnahmen des deutschen Ethikrates, der Bundesärztekammer et cetera, die alle in diese Richtung zielen.

Es gibt in der Tat eine Fülle ethischer Richtlinien für die KI-Forschung und -Entwicklung, doch klafft eine erhebliche Lücke zwischen den hochgesteckten Prinzipien und ihrer praktischen Umsetzung. Meta-Studien haben fast hundert solche Richtlinien identifiziert, dennoch entstehen nach diesen Kriterien laufend „unethische KI-Anwendungen". Die meisten Rahmenwerke für vertrauenswürdige KI sind zu abstrakt und bieten kaum praktische Orientierung – über 75 Prozent enthalten nur allgemeine Prinzipien, und mehr als 80 Prozent liefern praktisch keine konkreten Handlungsempfehlungen für Forschung und Entwicklung. Sie zeigen auf, *was* gemacht werden soll – aber nicht *wie*.

Es existiert sogar ein regelrechter „Markt" für Trustworthy KI, der Entwicklern und der Industrie ein „Ethics Washing" und „Ethics Shopping" ermöglicht. Es wird sich schon eine genügend abstrakte Guideline finden lassen, die zum eigenen Produkt passt. Und falls nicht, kann man sich immer noch eine selbst stricken – das nennt man dann „Ethics Lobbying" oder auch „Regulatory Capture". Microsoft tut dies derzeit beispielsweise, indem es im „Trustworthy and Responsible AI Network (TRAIN)" vier Initiativen zur Entwicklung von KI-Richtlinien im Gesundheitswesen ins Leben gerufen hat – und dafür Experten, technische Unterstützung und finanzielle Mittel bereitstellt. Dies ermöglicht dem Unternehmen, Teststandards und Vorschriften mitzugestalten, mit denen die eigene Technologie bevorzugt geprüft und der Markteintritt für Wettbewerber erschwert wird.

Vergessen wird bei alledem auch gerne, dass die meiste KI-„Forschung" in Wirklichkeit mehr Ingenieurwesen als wissenschaftliche Forschung ist. Der Schwerpunkt der aktuellen KI-Forschung im Gesundheitsbereich liegt hauptsächlich auf der Erkundung von Lösungen und Anwendungen sowie auf Machbarkeitsstudien (Proof of Concept, PoCs), die nicht ausreichend in der realen Welt validiert sind. Davon gibt es mehr als genug, es werden immer mehr – und meist werden sie uns auch als mehr verkauft: als einsatzreifes, „transformatives" Tool. Echte Validierungen von KI-Tools im Sinne qualitativ hochwertiger randomisiert kontrollierter Studien kann man an zwei Händen abzählen.

Diese Differenzierung zwischen Exploration beziehungsweise PoC auf der einen sowie Bestätigung und Validierung auf der anderen Seite wird derzeit weder bei der Bewertung von Studienergebnissen noch in der Diskussion über die Vertrauenswürdigkeit von KI ausreichend berücksichtigt. Dabei sollten unterschiedliche Vertrauenswürdigkeitsstufen für die explorative Phase und PoC beziehungsweise für die Validierung oder Implementierung gelten. Anforderungen an Art und Umfang der Trainingsdaten, an Erklärbarkeit und Transparenz, an Methoden zur Reduzierung von Verzerrungen (Bias) sowie an viele weitere Aspekte müssten je nach Entwicklungsstadium unterschiedlich ausgestaltet werden. Und das sollte dann auch entscheidend dafür sein, ob ein KI-Tool schon auf Patienten in der medizinischen Routine losgelassen werden darf.

Die eingangs erwähnte Arbeit, in der angeblich die Überlegenheit von ChatGPT gegenüber Ärzten in der Diagnosestellung gezeigt wurde, ist nicht nur ein Beispiel für den medialen Hype um KI, sondern auch für die mangelnde Qualität der Wissenschaft in diesem Feld. Die Stichprobengröße dieser Studie war 6, auf die auch dann noch die falsche Statistik angewendet wurde! Der 2018 verstorbene, bekannte britische Statistiker Douglas Altman pflegte zu sagen: „n=8 ist eine Dinnerparty, keine Studie". Zudem ging es in der Studie auch gar nicht darum, ob richtige Diagnosen gestellt wurden, sondern um diagnostisches Schlussfolgern („Reasoning"), das mit einer arbiträren, nicht-validierten Skala vermessen wurde. Dies nur eine Auswahl einer Vielzahl von Problemen, weswegen diese Arbeit niemals in *The Lancet – Digital Health* hätte publiziert werden dürfen. (Das übrigens am Rande auch zur verbreiteten Überschätzung der Qualitätskontrolle durch „Peer Review", über die sich der Narr schon häufiger kritisch geäußert hat – zum Beispiel in *LJ* 10/2020: 30-32).

»Die Wissenschaft zeigt erhebliche Defizite bei der Etablierung einer evidenzbasierten KI.«

Dass es auch anders geht, zeigt eine gut gemachte und frisch veröffentlichte Arbeit zur Genauigkeit von Mammographie-Screenings bei 105.000 (!) Frauen (*The Lancet Digital Health* 7(3): e175-e183). Die Ergebnisse der einfach verblindeten randomisiert kontrollierten Studie legen nahe, dass KI zur frühen Erkennung von klinisch relevantem Brustkrebs beitragen kann und die Arbeitsbelastung beim Screening reduziert, ohne die Anzahl der falsch-positiven Befunde zu erhöhen.

Selbstverständlich hat KI das Potenzial, die medizinische Diagnostik, klinische Entscheidungsfindung und Prognostik zu verbessern oder die Medikamentenentwicklung zu beschleunigen und tragbare Gesundheitstechnologien („Wearables") voranzutreiben – um nur einige der bekanntesten und häufig zitierten Anwendungsbereiche zu nennen. Allerdings wird die Entwicklung medizinischer KI-Anwendungen, die effektiv, sicher, vertrauenswürdig, fair und nachhaltig sind, derzeit durch eine „Move fast, break things"-Mentalität kommerzieller Entwickler sowie durch intensives Lobbying für eine industriefreundliche Regulierung behindert. Gleichzeitig zeigt die Wissenschaft – geblendet von dem Hype – erhebliche Defizite in Transparenz und Reporting sowie der Reduktion von Bias und einer kompetenten Validierung.

Kurzum: Wir brauchen eine evidenzbasierte KI.

Der Narr dankt Vince Madai für anregende Diskussionen und Kritik. Weiterführende Literatur und Links finden sich unter: dirnagl.com/lj.