Übersichtsarbeit

Qualität und Nutzen künstlicher Intelligenz in der Patientenversorgung

Kai Wehkamp, Michael Krawczak, Stefan Schreiber

Klinik für Innere Medizin I, Universitätsklinikum Schleswig-Holstein, Kiel: Prof. Dr. med. Kai Wehkamp, Prof. Dr. med. Dr. h.c. Stefan Schreiber

Department für Medizinmanagement, MSH Medical School Hamburg, Hamburg: Prof. Dr. med. Kai Wehkamp

Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität, Universitätsklinikum Schleswig-Holstein, Kiel: Prof. Dr. med. Michael Krawczak

Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität, Kiel: Prof. Dr. med. Dr. h.c. Stefan Schreiber

Zusammenfassung

Hintergrund: Künstliche Intelligenz (KI) wird zunehmend auch in der Patientenversorgung angewendet. Neben dem Wissen zur grundsätzlichen Funktionsweise dieser Verfahren wird es für Medizinerinnen und Mediziner künftig wichtig sein, Kenntnisse über Qualität, Nutzen und mögliche Risiken von KI-Anwendungen zu erlangen.

Methode: Diese Arbeit basiert auf einer selektiven Literaturrecherche zu Grundlagen, Qualität, Limitationen, Nutzen und Beispielen von KI-Anwendungen in der Patientenversorgung.

Ergebnisse: Es gibt eine wachsende Zahl von KI-Anwendungen in der Patientenversorgung (mehr als 500 Zulassungen in den USA). Ihre Qualität und ihr Nutzen fußen auf verschiedenen, wechselseitig voneinander abhängigen Faktoren. Diese umfassen die reale Lebenswelt, die Art und den Umfang der darin erhobenen Daten, die Auswahl der von der KI verwendeten Variablen, die genutzten Algorithmen sowie das Ziel und den Einsatz der jeweiligen Anwendung. Auf all diesen Ebenen kann es zu (versteckten) Verzerrungen und Fehlern kommen. Bei der Bewertung von Qualität und Nutzen einer KI-Anwendung müssen daher die wissenschaftlichen Prinzipien der evidenzbasierten Medizin berücksichtigt werden – eine Forderung, die nicht immer transparent umgesetzt wird.

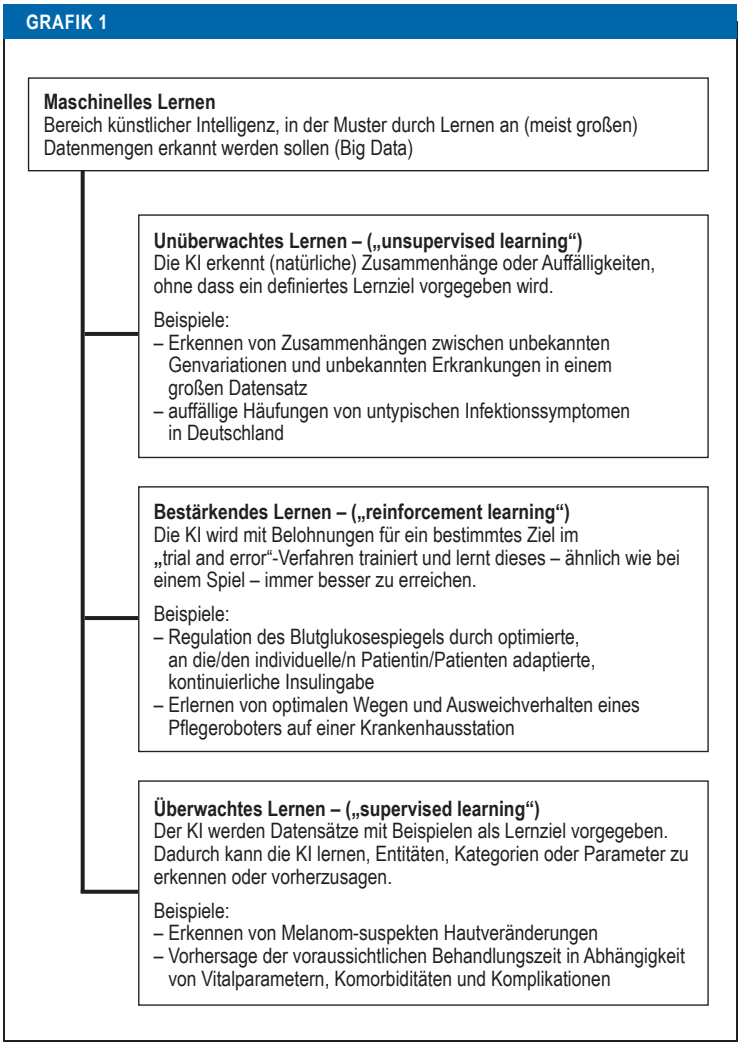
Schlussfolgerung: KI hat das Potenzial, die Patientenversorgung zu verbessern und dabei den Herausforderungen einer stetig wachsenden Informations- und Datenflut in der Medizin bei gleichzeitig begrenzten Personalressourcen zu begegnen. Im Zuge dessen müssen jedoch die Limitationen und Risiken von KI-Anwendungen kritisch und verantwortungsvoll reflektiert werden. Wichtige Grundlage hierfür ist neben wissenschaftlicher Transparenz die Stärkung der fachlichen Kompetenz der Ärztinnen und Ärzte.

Zitierweise
Wehkamp K, Krawczak M, Schreiber S:
The quality and utility of artificial intelligence in patient care. Dtsch Arztebl Int 2023; 120: 463–9.
DOI: 10.3238/arztebl.m2023.0124

Menschliche Intelligenz ist eines der bemerkenswertesten Ergebnisse der Evolution. Von entscheidender Bedeutung für die Intelligenzleistung unseres Gehirns ist seine Fähigkeit, Modelle zu bilden, die ein detailliertes Abbild der komplexen Realität liefern mit dem Ziel, Vorhersagen im Dienst einer erfolgreichen Interaktion mit unserer Umwelt zu treffen (1). Künstliche Intelligenz (KI) ist demgegenüber ein Sammelbegriff für Verfahren, die es Computern erlauben, Aufgaben zu erfüllen, die normalerweise menschliche Intelligenz erfordern. In diesem Sinne handelt es sich bereits bei den Algorithmen eines einfachen Schachcomputers um eine KI. Eine Form der KI ist das sogenannte maschinelle Lernen (ML, „machine learning“), bei dem Muster aus Daten abgeleitet werden, um entweder die zugrunde liegenden Daten besser zu interpretieren oder auf ihrer Grundlage bestimmte Vorhersagen zu treffen.

Im Bereich des ML gab es in den letzten zehn Jahren bedeutende Fortschritte, unter anderem durch die Entwicklung mehrschichtiger („tiefer“) künstlicher neuronaler Netze (DNN, „deep neuronal network) (2). Allerdings werden wohl noch Jahre bis Jahrzehnte vergehen, ehe ML beziehungsweise KI dem breiten Spektrum menschlicher Intelligenz vollumfänglich gleichkommt (falls dies überhaupt jemals möglich sein wird) (3). Dessen ungeachtet erzielt KI in Form des ML bereits heute in Teilen der Medizin Resultate, die die menschliche Leistungsfähigkeit übertreffen. Die künf-

GRAFIK 1



Prinzipielle Konzepte des maschinellen Lernens

tige Entwicklung solcher Verfahren muss jedoch kritisch und mit unabhängigem Sachverstand begleitet werden, damit die Medizin auch weiterhin der Maxime einer bestmöglichen Patientenversorgung gerecht werden kann. Hierbei steht die Ärzteschaft in einer besonderen Verantwortung.

Diese Arbeit gibt einen Überblick über wichtige Aspekte der Beurteilung von Qualität, Nutzen und Limitationen von KI-Anwendungen in der Patientenversorgung, auch um damit einen Beitrag zum verantwortungsvollen Einsatz dieser Technik zu leisten.

Methode

Basierend auf einer selektiven Literaturrecherche in PubMed werden ausgewählte Aspekte der Beurteilung von Qualität und Nutzen von (insbesondere ML-basierten) KI-Anwendungen in der Patientenversorgung dargestellt. Diese Darstellung des Status quo wird exemplarisch um aktuelle Anwendungsbeispiele ergänzt, die einschlägigen Fachmedien und wissenschaftlichen Studien entnommen wurden.

Ergebnisse

Daten als Grundlage des maschinellen Lernens

Maschinelles Lernen (ML) basiert auf Daten als beispielhafte Repräsentation einer bestimmten Lernwelt. In den Lerndaten sollen Muster oder abstrakte Regeln erkannt und anschließend auf neue Daten angewandt werden, um Charakteristika zu erkennen, vorherzusagen oder Aussagen zu generieren. ML weist damit konzeptuell eine große Ähnlichkeit zum menschlichen Lernen aus Beispielen und dem Erkennen von Ähnlichkeiten und Unterschieden auf.

Je unstrukturierter die verwendeten Daten sind, und je mehr unterschiedliche Datenmodalitäten zusammengefasst werden sollen, desto höher sind die Herausforderungen an eine KI-Anwendung (4). So gibt es zwar KI-basierte Verfahren, die Brustkrebs in Mammografie-Bildern mit einer Sensitivität und Spezifität erkennen, die der eines durchschnittlich versierten Radiologen (aber bislang nicht eines Experten) vergleichbar sind (5). Der KI-basierte Erkenntnisgewinn aus einer Kombination von verschiedenen, unstrukturierten Datentypen wie etwa DNA-Sequenzen, histopathologische Bilder und Laborbefunde misslingt in der Praxis jedoch noch (6). Außerdem birgt die Nutzung umfangreicher medizinischer Daten stets die Gefahr einer Verletzung individueller Persönlichkeitsrechte, woraus datenschutzrechtliche Einschränkungen resultieren können (7). Aktuell zählen die limitierte Qualität und Verfügbarkeit der komplexen und heterogenen Daten zu den in weiten Bereichen nicht zufriedenstellend lösbaren Herausforderungen für den Einsatz von medizinischen KI-Anwendungen.

Konzepte des maschinellen Lernens in der Medizin

ML-Ansätze lassen sich primär in drei Gruppen unterteilen *(Grafik 1)*:

- Das unüberwachte Lernen versucht, ohne konkrete Vorgaben Zusammenhänge, Strukturen oder Anomalien in Daten zu identifizieren. Dieser Ansatz wird zum Beispiel zur Identifikation von Subgruppen in Multiomics-Datensätzen verwendet (8). In der Patientenversorgung befinden sich Verfahren des unüberwachten Lernens noch in einem experimentellen Stadium, perspektivisch ist deren Verwendung aber zum Beispiel bei der syndromalen Überwachung denkbar – etwa im Rahmen eines Outbreak-Monitorings für Infektionserkrankungen (9).
- Das bestärkende Lernen trainiert im Hinblick auf Belohnungen, die für ein bestimmtes Outcome vergeben werden. Auch dieser Ansatz wird in der Medizin bislang nur in Studien untersucht, könnte künftig aber geeignet sein, um zum Beispiel in Closed-Loop-Verfahren die Insulingabe an die/den individuelle/n Patientin/Patienten anzupassen (10).
- Ansätze des überwachten Lernens zielen häufig auf das Klassifizieren von Daten oder die Prädiktion künftiger Ereignisse ab. Die entsprechenden Algorithmen werden mit Trainingsdaten angelernt, in denen das Lernziel vorgegeben ist (zum Beispiel Röntgenbilder mit markierten Raumforderungen und Vergleichsbilder,

die keine Raumforderung enthalten). Die erkannten Muster werden dann hinsichtlich ihrer Güte an Testdatensätzen validiert. Die meisten bereits in Zulassung befindlichen KI-Anwendungen basieren auf überwachtem Lernen aus einheitlichen, monomodalen Daten (zum Beispiel der alleinigen Verarbeitung von Bildern möglicher Hautveränderungen zur Erkennung bösartiger Veränderungen) (11, 12).

Risiken und Limitationen von KI-Anwendungen

Um Risiken und Limitationen der Aussagekraft von ML-Anwendungen beurteilen zu können, ist eine Kenntnis des stets zugrunde liegenden ML-Zyklus bedeutsam, der auf mehreren, stark voneinander abhängigen Ebenen basiert *(Grafik 2)*. An erster Stelle stehen dabei die Gegebenheiten der realen Lebenswelt, die möglichst repräsentativ in Form digitaler Daten abgebildet werden. Die zugehörigen Variablen müssen ausgewählt und vorbereitet werden (sogenanntes „feature selection and engineering", bei DNNs teilweise entfallend), um dann vom ML-Algorithmus verarbeitet zu werden. Die Ergebnisse werden von Anwendern (das heißt Medizinerinnen und Medizinern) genutzt und entfalten damit wiederum Wirkung auf die reale Lebenswelt (das heißt die Versorgung von Patientinnen und Patienten) (13).

Auf jeder Ebene des ML-Zyklus wirkt eine Vielzahl teils redundanter Einflussfaktoren, die die Ergebnisse einer KI-Anwendung erheblich verzerren und ihre Validität limitieren können. Diese Limitationen sind maßgeblich dafür verantwortlich, dass die praktische Anwendung von KI in der Patientenversorgung in vielen Bereichen noch immer hinter den Erwartungen und Hoffnungen zurückbleibt. Eine kritische Reflexion der einzelnen Ebenen des ML-Zyklus ist deswegen essenziell für die realistische Bewertung der Potenziale und Qualitäten von ML-Anwendungen (14, 15).

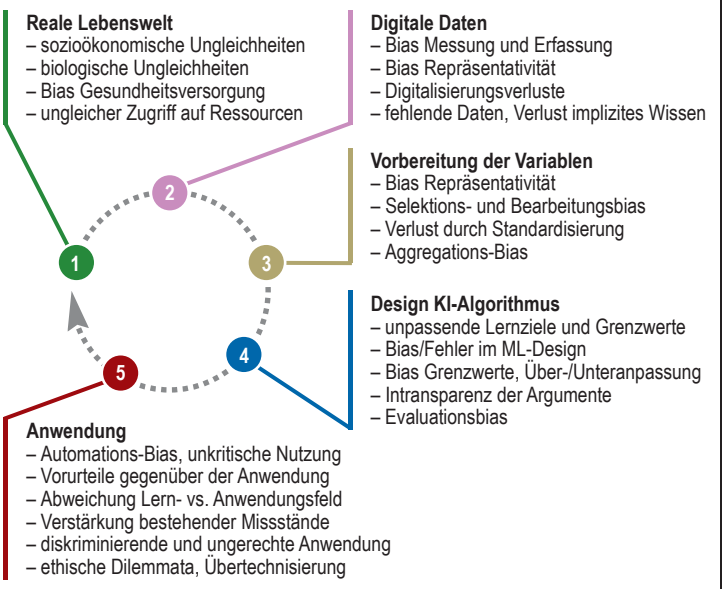
Reale Lebenswelt

Menschen leben in realen Lebenswelten. Diese sind regelhaft von sozioökonomischen, biologischen und anderen Inhomogenitäten gekennzeichnet, die mit einer Gefährdung oder Benachteiligung bestimmter Individuen oder Bevölkerungsgruppen einhergehen können. Bei der Erhebung der einer KI-Anwendung zugrunde liegenden Daten sind solche potenziellen Verzerrungen zu berücksichtigen und gegebenenfalls auszugleichen (14, 16, 17).

Digitale Daten

Daten können die reale Lebenswelt grundsätzlich nur unvollständig und in Teilaspekten repräsentieren. Um trotzdem ein hinreichend gutes Abbild der realen Lebenswelt zu erlangen, muss die Datenerhebung selbst so objektiv, präzise und genau wie möglich sein. Zudem ist bei der Auswahl der Datenquellen auf eine angemessene Repräsentativität zu achten (18). Viele, insbesondere Individuum-spezifische medizinische Informationen lassen sich jedoch nur in Textform mithilfe der Komplexität natürlicher Sprache erfassen, also in Form unstrukturierter Daten, die durch Spracherkennung vorbearbeitet werden müssen (Natural Language Processing) (19). Und schließlich können Informationen, die sich nicht digital dokumentieren lassen, für KI-Anwendungen generell nicht nutzbar gemacht werden (so zum Beispiel die von Erfahrung und Intuition geprägte Einschätzung des Gesamtbildes einer Patientin/eines Patienten) (15, 20).

GRAFIK 2



Ausgewählte Limitationen und Risken für die Qualität von Anwendungen künstlicher Intelligenz (KI) auf Ebenen des Lern- und Anwendungszyklus des maschinellen Lernens (ML)

Auswahl und Vorbereitung der Variablen

Um mittels KI-Anwendungen möglichst valide Modelle der realen Lebenswelt zu erhalten, müssen die darin einbezogenen Variablen passend gewählt und vorbereitet werden (zum Beispiel Beschränkung auf Röntgen-Befunde und spezifische klinische Parameter in der onkologischen Diagnostik). Diese Auswahl sowie die spätere Standardisierung und Normalisierung der Daten können deren Repräsentativität einschränken und die Validität der Ergebnisse einer KI-Anwendung limitieren (15).

Design der Algorithmen

Das Design eines ML-Algorithmus umfasst die Programmierung des Software-Codes und die Integration der zuvor ausgewählten Variablen. Auch auf dieser Ebene kann es zu Fehlern und Verzerrungen kommen, etwa durch die mangelhafte Berücksichtigung von Besonderheiten der zu verwendenden Daten, unscharfe Zieldefinitionen für die Mustererkennung oder die Einbettung unpassender Grenzwerte (13, 21). Um eine hinreichende Akzeptanz und kritische Reflexion des Designs durch die Anwenderinnen und Anwender sicher-

TABELLE

Beispiele zugelassener (oder sich im Zulassungsprozess befindlicher) KI-basierter Anwendungen in der Patientenversorgung*

Fachbereich/Funktion	Datenbasis	Beispiel Anwendung	KI-Konzept	assoziierte Studien	Ergebnisse
Dermatologie Erkennung dermales Melanom	dermatoskopische Bilder	Moleanalyzer pro	überwachtes Lernen	Haenssle et al., Ann Oncol, 2020 (e1)	Sensitivität: 95 % (vs. Expertin/Experte: 89 %) Spezifität: 76,7 % (vs. Expertin/Experte:80,7 %) Treffergenauigkeit: 84 % (vs. Expertin/Experte: 84 %)
Diabetologie Steuerung Blutglukosespiegel auf Intensivstation	Glukosespiegel, Kohlenhydrataufnahme, Insulingabe	Space GlucoseControl System	Algorithmus zur modelbasierten Prädiktion	Blaha et al, BMC Anesthesiology, 2016 (e2)	Blutglukosespiegel im Zielbereich: 83 % Zeit Episoden schwerer Hypoglykämien: 0,01 % Zeit
Gastroenterologie Erkennung kolorektale Neoplasie	Koloskopie Bilder	GI-Genius	überwachtes Lernen	Repici et al., Gastroenterology, 2020 (e3)	Sensitivität: 99,7 % Spezifizität: 91,1 % Detektionsrate mit KI: 54,8 % (vs. 40,4 % ohne KI-Unterstützung)
Gynäkologie Erkennung Mammakarzinom	digitale Mammographie	Transpara	überwachtes Lernen	Romere-Martín et al., Radiology, 2022 (e4)	Sensitivität: 70,8 % (vs. zweifache Befundung Expertin/Experte: 67,3 %)
Herzchirurgie Vorhersage postoperatives Risiko Nachblutung nach Herz-OP	strukturierte klinische Daten (Vitalparameter, OP, Testergebnisse, etc.)	x-c-bleeding	überwachtes Lernen	Meyer et al., Lancet Respiratory Medicine, 2018 (e5)	Sensitivität: 74 % (vs. Bojar-Algorithmus: 21 %) Spezifität: 84 % (vs. Bojar-Algorithmus: 95 %) Treffergenauigkeit: 80 % (vs. Bojar-Algorithmus: 58 %) PPV 84 % (vs. Bojar-Algorithmus: 81 %)
Kardiologie Erkennung Myokardischämie	vektorkardiografische Daten	Cardisio	überwachtes Lernen	Braun et al., Journal of Electrocardiology, 2020 (e6)	Sensitivität: 90,2 %/97,2 % (Frauen/Männer) Spezifität: 74,4 %/76,1 % (Frauen/Männer) Treffergenauigkeit: 82,5 %/90,7 % (Frauen/Männer)
Nephrologie Vorhersage postoperatives Risiko Nierenschädigung nach Herz-OP	strukturierte klinische Daten (Vitalparameter, OP, Testergebnisse, etc.)	x-c-renal injury	überwachtes Lernen	Meyer et al., Lancet Respiratory Medicine, 2018 (e5)	Sensitivität 94 % (vs. KDIGO Nierenversagen: 53 %) Spezifität: 86 % (vs. KDIGO Nierenversagen 92 %) Treffergenauigkeit: 90 % (vs. KDIGO Nierenversagen 73 %) PPV: 87 % (vs. KDIGO Nierenversagen 87 %)
Ophthalmologie Erkennung diabetische Retinopathie (mtmDR)	undilatierte 2-Feld-Fundusfotografie	EyeArt	überwachtes Lernen	Ipp et al., JAMA network open, 2021 (e7)	Sensitivität: 95,5 % Spezifizität: 85,0 % PPV: 59,5 %
Orthopädie Optimierung Funktionalität Oberarmprothese	kontinulieriche elektromyografische Ableitung	Myo Plus	algorithmenbasierte Mustererkennung	Franzke et al., Plos One, 2019 (e8)	qualitative Beurteilung: intuitive Kontrolle möglich, im täglichen Gebrauch nicht immer zuverlässig, hoher Trainingsbedarf (besser als konventionelle Kontrolle)
Pathologie Detektion NSCLC -Tumor-Zellen (Lunge)	Histopathologie mit immunhistochemischer Färbung für PD-L1	Mindpeak Lung (NSCLC) PD-L1	überwachtes Lernen	Daifalle, Günther, Mindpeak Website, 2022 (e9)	Übereinstimmungsrate zur Expertinnen-/Expertenbeurteilung (bei 1 % Anteil Tumorzellen in Analyse): 85 % (vs. konventionelle Befundung: 83 %)
Psychiatrie Früherkennung Risiko Delir bei stationären Patienten	strukturierte klinische Daten (Vitalparameter, Laborergebnisse, Diagnosen etc.)	Clinalytix	überwachtes Lernen	Sun et al, Journal of Medical Internet Research, 2022 (e10)	Sensitivität: 80 % Spezifität: > 85 %
Radiologie Detektion Lungenrundherd	Low-Dose-CT-Thorax	AI-Rad Companion Chest CT	überwachtes Lernen	Chamberlin et al., BMC Medicine, 2021 (e11)	Sensitivität: 100 % Spezifizität: 70 % PPV: 83,1 % Variabilität KI vs. Expertin/Experte: 0,741 (Cohen's Kappa)

* Dargestellt sind die Ergebnisse assoziierter Studien; weitere Kennzahlen, Referenzverfahren und Studienkonzepte finden sich in den jeweiligen Publikationen (Referenzen siehe Anhang).
Die Anwendungen bieten in der Regel nur eine Entscheidungsunterstützung und dürfen die ärztliche Entscheidung nicht ersetzen. Die Darstellung zeigt nur einen exemplarischen Ausschnitt zugelassener Anwendungen.
CT, Comptertomografie; KI, künstliche Intelligenz; mtmDR, „more than mild diabetic retinopathy"; KDIGO, Kidney Disease: Improving Global Outcomes; NSCLC, nichtkleinzelliges Bronchialkarzinom oder Lungenkarzinom;
PPV, positiver prädiktiver Wert

zustellen, sollte der Algorithmus zudem Erklärungen für die jeweils erzielten Ergebnisse liefern können (sogenannte Explainability) (22, 23).

Anwendung in der realen Lebenswelt

Auf der Ebene der praktischen Anwendung, also der Patientenversorgung, können sich Fehler und Verzerrungen aller vorhergehenden Ebenen des ML-Zyklus negativ auswirken. Besondere Gefahren entstehen dabei durch nicht berücksichtigte Unterschiede zwischen Lern- und Anwendungswelt und durch die mangelnde Ausrichtung von KI-Anwendungen am späteren praktischen Einsatz (24). Unpräzise Ergebnisse, technische Hürden, mangelnde inhaltliche Transparenz und Misstrauen führen schnell dazu, dass das Potenzial von KI-Anwendungen für die Patientenversorgung nicht voll ausgeschöpft wird, zum Beispiel wenn sich Programme zur KI-basierten Analyse von histopathologischen Befunden nicht in die bestehenden Abläufe integrieren lassen oder diese nicht zu einer Zeitersparnis führen (18, 25, 26). Auf der anderen Seite kann die unkritische, zu vertrauensvolle Nutzung von KI-Anwendung dazu führen, dass zum Beispiel wichtige differenzialdiagnostische Überlegungen in der Praxis ausgeklammert werden. Prinzipiell birgt das unreflektierte Verfolgen KI-basierter Behandlungskonzepte die Gefahr, dass die Medizin durch eine Übertechnisierung wichtiger menschlicher Faktoren beraubt wird. So stellt etwa die quasi-objektivierende Berechnung von Outcome-Wahrscheinlichkeiten eine große Herausforderung für die differenzierte Kommunikation zwischen Ärztin/Arzt und Patientin/Patient dar (27, 28). Auch können sich ethische Dilemmata zuspitzen, wenn die Ergebnisse von KI-Anwendungen unreflektiert als Grundlage für Allokations- und Priorisierungsentscheidungen genutzt werden (29).

Qualität und Nutzen klinischer KI-Anwendungen

Evidenzgrundlage für die Beurteilung von ML-Anwendungen

Eine wissenschaftliche Grundlage gehört zu den wesentlichen Qualitätsansprüchen der modernen Medizin. Entsprechend sollte auch für KI-basierte medizinische Anwendungen deren Objektivität (Unabhängigkeit von unkontrollierten Einflussfaktoren), Reliabilität (Verlässlichkeit) und Validität (Gültigkeit) transparent beurteilbar sein. Um die Güte von entscheidungsunterstützenden KI-Algorithmen darzustellen, werden meistens die statistischen Testgrößen Sensitivität, Spezifität und Präzision (positiv prädiktiver Wert) verwendet. Diese sollten um eine kritische Beurteilung von Bias und Risiken im jeweiligen ML-Zyklus ergänzt werden *(Grafik 2)*. Darüber hinaus gehört zu einer evidenzbasierten Nutzenbewertung die Untersuchung der Methode im realen Setting und im Vergleich zu alternativen Verfahren, analog der Vorgehensweise einer klinischen Studie (zum Beispiel als prospektive Interventionsstudie, die ein KI-basiertes mit einem klassischen Diagnostikverfahren vergleicht) (30).

Je nach Anwendung sollten neben der Genauigkeit auch passende patientenbezogene Endpunkte wie Lebensqualität, Überlebenszeit, Krankheitsprogress und Symptomreduktion bewertet werden. Idealerweise erfolgt sogar bereits das Training einer KI-Anwendung mit Blick auf die Verbesserung solcher patientenbezogenen Endpunkte. Nur so kann die Anwendung perspektivisch besser werden als eine von Menschen vorgenommene Bewertung von Diagnose- und Behandlungsdaten (24). Bislang gibt es jedoch nur wenige prospektive Studien, die KI-Anwendungen im Vergleich zum Status quo der medizinischen Versorgung untersuchen oder diesbezüglich bereits einen Nutzen darstellen konnten (31, 32). Eine umfassende Bewertung des Mehrwerts einer KI-Anwendung schließt neben der Beurteilung potenzieller Risiken für die Patientensicherheit auch die der Wirtschaftlichkeit (inklusive möglicher Zeit- und Ressourcenersparnisse) und der ethischen und soziokulturellen Folgen ein (33–36).

Praktischer Einsatz von ML-basierten Anwendungen in der Patientenversorgung

In den USA listet die Food and Drug Administration (FDA) aktuell 521 zugelassene medizinische KI-Anwendungen auf (37). In Deutschland fehlen offizielle Angaben; es ist bislang von einigen Dutzend Zulassungen auszugehen.

Gemessen an der Zahl KI-bezogener Publikationen ist die tatsächliche Bedeutung von in Deutschland zugelassenen Anwendungen eher überschaubar. Grund dafür sind insbesondere die genannten Limitationen in Bezug auf die Datengrundlage, die eingeschränkte Übertragbarkeit zwischen Lern- und Anwendungswelt sowie Herausforderungen hinsichtlich einer praktikablen und ökonomisch sinnvollen Einbindung in bestehende Versorgungsprozesse (25). KI-Anwendungen für die Patientenversorgung sind in der Regel Medizinprodukte, die nur nach Durchführung einer Konformitätsbewertung für die jeweilige Risikoklasse vertrieben beziehungsweise eingesetzt werden dürfen. Ihre Zulassung bezieht sich dabei formal in der Regel nur auf die Entscheidungsunterstützung und setzt voraus, dass die Verantwortung bei den sie einsetzenden Medizinerinnen und Medizinern verbleibt. Ein unreflektierter Einsatz der Verfahren (zum Beispiel im Sinne eines Automatisierungs-Bias) kann also Risiken beinhalten (18). Außerdem ist es für die Zulassung von Medizinprodukten bislang nicht immer erforderlich, nutzenorientierte Anwendungsstudien zu veröffentlichen. Vielmehr erfolgt oft ein nicht wissenschaftlich zu veröffentlichender Nachweis der Funktionalität im Rahmen der Zulassung oder die zugehörigen Studien fanden in einem artifiziellen Setting statt. Entsprechend intransparent ist die Darstellung von Qualität und Nutzen vieler der in Deutschland bereits im Einsatz befindlichen ML-basierten Anwendungen.

In der *Tabelle* findet sich eine synoptische Darstellung einiger Beispiele von in Deutschland zugelassenen (oder im Zulassungsprozess befindlichen) Systemen mit einer exemplarischen Nennung der jeweils verfügbaren wissenschaftlichen Evidenz. Der Großteil der Anwen-

dungen beruht auf monomodalen, einheitlichen Daten. Insgesamt ergibt sich hinsichtlich der publikatorischen Grundlage zugelassener KI-Anwendungen ein uneinheitliches und bisweilen intransparentes Bild. Für einige Verfahren lässt sich der Nutzen aus publizierten Anwendungsstudien entnehmen, zum Beispiel für die ML-basierte Unterstützung der koloskopischen Detektion kolorektaler Polypen oder die fotobasierte Erkennung maligner Hautveränderungen. Die Leistungsfähigkeit dieser Anwendungen hat sich in beiden Fällen als vergleichbar mit einem Standardverfahren erwiesen (12, 38). Für andere zugelassene Anwendungen wurden entweder keine klinischen Nutzendaten oder lediglich ausgewählte statistische Kenngrößen veröffentlicht (zum Beispiel Sensitivität, Spezifität). Bei anderen Anwendungen lässt sich der Nutzen nicht über einen klinischen Zusatznutzen, sondern durch effizientere Prozesse oder die Senkung von Versorgungshürden darstellen. Beispiele hierfür liefern Bereiche, in denen spezifisches Fachwissen vor Ort fehlt (zum Beispiel bei der Erkennung seltener elektrokardiografischer Befunde) oder in denen ein hoher Durchsatz erforderlich ist (zum Beispiel für das Mammografie-Screening). Und schließlich können Kostenersparnisse und die Vereinfachung des Zugangs zu bestimmten Behandlungsverfahren in unterversorgten Regionen erheblich zum praktischen Nutzen von KI-Anwendungen beitragen, wie etwa bei der Diagnostik der diabetischen Retinopathie oder des malignen Melanoms (e1–e11).

Resümee

Ein rasant anwachsender Wissens- und Informationsstand und die daraus resultierenden, neuen diagnostischen und therapeutischen Möglichkeiten stellen die Medizin vor die Herausforderung, diese Informationen entweder so zu verdichten, dass sie handhabbar bleiben, oder sie in ihrer Gänze bestmöglich zum Wohle von Patientinnen und Patienten sowie der Gesellschaft zu nutzen.

Medizinerinnen und Mediziner müssen heute einen immer größeren Aufwand betreiben, um dem aktuellen Stand von Wissenschaft und Technik zu genügen und gleichzeitig den ökonomischen Rahmenbedingungen und den Ansprüchen an eine menschliche Medizin gerecht zu werden. Ihre Kapazitäten stoßen dabei teilweise an Grenzen. Maschinelles Lernen als aktuell wirkmächtigste Entwicklung der künstlichen Intelligenz ahmt das menschliche Lernen nach und kann in Abhängigkeit von Datenqualität und verfügbarer Rechenleistung immer bessere medizinische Vorhersagen und Klassifizierungen liefern.

Die aktuelle Studienlage zeigt, dass die präventive, diagnostische und therapeutische Patientenversorgung zunehmend von einer KI-Unterstützung profitieren kann. Allerdings muss jede Technik, die mittelbar Auswirkungen auf die medizinische Praxis hat und somit potenziell auf Leben, Krankheit und Tod von Menschen, mit besonderer Sorgfalt hinsichtlich Nutzen und Risiken geprüft werden. Im Lern- und Anwendungszyklus des ML kann es auf verschiedenen Ebenen zu Risiken durch mögliche Verzerrungen, negative Verstärkungen oder Fehler kommen. KI-Anwendungen stellen daher ein potenzielles Risiko für die Patientinnen und Patienten dar und müssen deswegen bislang noch kritisch durch menschliche Urteilskraft geprüft werden.

Für die Qualitätsbeurteilung von KI-Anwendungen sind breite Kompetenzen erforderlich, die von der originären medizinischen Expertise über die Gestaltung der Versorgungsprozesse, Datenwissenschaften, Informatik bis hin zu Ethik und Recht reichen. Gerade weil nicht alle diese Facetten zur engeren medizinischen Domäne gehören, muss sich die Ärzteschaft ein übergreifendes Verständnis der KI aneignen, um ihrer gesellschaftlichen Verantwortung für deren kritisch reflektierten Einsatz in der Patientenversorgung gerecht werden zu können (zum Beispiel über den Kurs: www.ki-campus.org/courses/drmedki_basics_cme) *(eKasten)* (39, 40).

Wegen der Komplexität und tendenziellen Intransparenz von KI-Anwendungen bedarf es außerdem regulativer Sicherungsmechanismen, die einen starken Fokus auf den praktischen Nutzen, die anwendungsbezogenen, teils erheblichen Risiken und die Sicherstellung einer hohen inhaltlichen Transparenz setzen. Verantwortungsvoll eingesetzt, kann KI künftig eine evidenzbasierte und wirtschaftliche Patientenversorgung fördern und gleichzeitig das menschliche Wesen (und die menschliche Intelligenz) der Medizin unterstützen.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt besteht.

Manuskriptdaten
eingereicht: 30.11.2022, revidierte Fassung angenommen: 08.05.2023

Literatur

1. Hawkins J, Lewis M, Klukas M, Purdy S, Ahmad S: A framework for intelligence and cortical function based on grid cells in the neocortex. Front Neural Circuits 2019; 12: 121.
2. Topol EJ: High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; 25: 44–56.
3. Katritsis DG: Artificial intelligence, superintelligence and intelligence. Arrhythm Electrophysiol Rev 2021; 10: 223–4.
4. Zhang D, Yin C, Zeng J, Yuan X, Zhang P: Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 2020; 20: 280.
5. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al.: Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019; 111: 916–22.
6. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ: Multimodal biomedical AI. Nat Med 2022; 28: 1773–84.
7. Vidalis T: Artificial intelligence in biomedicine: a legal insight. BioTech (Basel) 2021; 10: 15.
8. Eicher T, Kinnebrew G, Patt A, et al.: Metabolomics and multi-omics integration: a survey of computational methods and resources. Metabolites 2020; 10: 202.
9. Wen A, Wang L, He H, et al.: An aberration detection-based approach for sentinel syndromic surveillance of COVID-19 and other novel influenza-like illnesses. J Biomed Inform 2021; 113: 103660.
10. Tejedor M, Woldaregay AZ, Godtliebsen F: Reinforcement learning application in diabetes blood glucose control: a systematic review. Artif Intell Med 2020; 104: 101836.
11. Rajpurkar P, Chen E, Banerjee O, Topol EJ: AI in health and medicine. Nat Med 2022; 28: 31–8.
12. Haenssle HA, Fink C, Toberer F, et al.: Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020; 31: 137–43.

13. Kocak B, Kus EA, Kilickesmez O: How to read and review papers on machine learning and artificial intelligence in radiology: a survival guide to key methodological concepts. Eur Radiol 2021; 31: 1819–30.
14. Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A: Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? BMJ 2021; 372: n304.
15. Suresh H, Guttag J: A framework for understanding sources of harm throughout the machine learning life cycle. In: ACM International Conference Proceeding Series 2021. www.doi.org/10.1145/3465416.3483305 (last accessed on 16 March 2022).
16. Celi LA, Cellini J, Charpignon ML, et al.: Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. PLOS Digit Health 2022; 1: e0000022.
17. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z: An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021; 27: 136–140.
18. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K: Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019; 28: 231–7.
19. Goh KH, Wang L, Yeow AYK, et al.: Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021; 12: 711.
20. van der Niet AG, Bleakley A: Where medical education meets artificial intelligence: 'Does technology care?' Med Educ 2021; 55: 30–6.
21. Barboi C, Tzavelis A, Muhammad LN: Comparison of severity of illness scores and artificial intelligence models that are predictive of intensive care unit mortality: meta-analysis and review of the literature. JMIR Med Inform 2022; 10: e35293.
22. Loftus TJ, Tighe PJ, Ozrazgat-Baslanti T, et al.: Ideal algorithms in healthcare: explainable, dynamic, precise, autonomous, fair, and reproducible. PLOS Digit Health 2022; 1: e0000006.
23. Amann J, Vetter D, Blomberg SN, et al.: To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health 2022; 1: e0000016.
24. Obermeyer Z, Topol EJ: Artificial intelligence, bias, and patients' perspectives. Lancet 2021; 397(10289): 2038.
25. Cabitza F, Campagner A, Balsano C: Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med 2020; 8: 501.
26. Gaube S, Suresh H, Raue M, et al.: Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021; 4: 31.
27. Nagy M, Sisk B: How will artificial intelligence affect patient-clinician relationships? AMA J Ethics 2020; 22: E395–400.
28. Lu SC, Xu C, Nguyen CH, Geng Y, Pfob A, Sidey-Gibbons C: Machine learning-based short-term mortality prediction models for patients with cancer using electronic health record data: systematic review and critical appraisal. JMIR Med Inform 2022; 10: e33182.
29. Wingfield LR, Ceresa C, Thorogood S, Fleuriot J, Knight S: Using artificial intelligence for predicting survival of individual grafts in liver transplantation: a systematic review. Liver Transpl 2020; 26: 922–34.
30. Caliebe A, Leverkus F, Antes G, Krawczak M: Does big data require a methodological change in medical research? BMC Med Res Methodol 2019; 19: 125.
31. Nagendran M, Chen Y, Lovejoy CA, et al.: Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020; 368: m689.
32. Zhou Q, Chen ZH, Cao YH, Peng S: Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med 2021; 4: 154.
33. Ryan M: In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 2020; 26: 2749–67.
34. Collins GS, Dhiman P, Andaur Navarro CL, et al.: Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021; 11: e048008.
35. Wiens J, Saria S, Sendak M, et al.: Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019; 25: 1337–40.
36. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K: The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019; 25: 30–6.
37. FDA US: Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (last accessed on 5 May 2023).
38. Repici A, Spadaccini M, Antonelli G, et al.: Artificial intelligence and colonoscopy experience: lessons from two randomised trials. Gut 2022; 71: 757–65.
39. Keane PA, Topol EJ: AI-facilitated health care requires education of clinicians. Lancet 2021; 397 (10281): 1254.
40. Young AT, Amara D, Bhattacharya A, Wei ML: Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digital Health 2021; 3: e599–e611.

Anschrift für die Verfasser
Prof. Dr. med. Kai Wehkamp, MPH
Klinik für Innere Medizin I
Universitätsklinikum Schleswig-Holstein
Arnold-Heller-Straße 3 (Haus 6), 24105 Kiel
Kai.Wehkamp@uksh.de

Zitierweise
Wehkamp K, Krawczak M, Schreiber S: The quality and utility of artificial intelligence in patient care. Dtsch Arztebl Int 2023; 120: 463–9. DOI: 10.3238/arztebl.m2023.0124

► **Die englische Version des Artikels ist online abrufbar unter: www.aerzteblatt-international.de**

Zusatzmaterial
eLiteratur, eKasten:
www.aerzteblatt.de/m2023.0124 oder über QR-Code

Hinweise für Autoren von Diskussionsbeiträgen im Deutschen Ärzteblatt

- Reichen Sie uns bitte Ihren Diskussionsbeitrag bis spätestens vier Wochen nach Erscheinen des Primärartikels ein.
- Argumentieren Sie wissenschaftlich, sachlich und konstruktiv. Briefe mit persönlichen Angriffen können wir nicht abdrucken.
- Schreiben Sie klar und deutlich, fokussieren Sie sich inhaltlich. Vermeiden Sie es, Nebenaspekte zu berühren.
- Sichern Sie die wichtigsten Behauptungen durch Referenzen ab. Bitte geben Sie aber – abgesehen von dem Artikel, auf den Sie sich beziehen – insgesamt nicht mehr als drei Referenzen an.
- Beschränken Sie Ihren Diskussionsbeitrag auf eine Textlänge von 250 Wörtern bei Zuschriften zu Original- oder Übersichtsarbeiten und auf 150 Wörter bei Kommentaren zu Kurzmitteilungen oder Klinischen Schnappschüssen (ohne Referenzen und Autorenadresse).
- Verzichten Sie auf Tabellen, Grafiken und Abbildungen. Aus Platzgründen können wir solche grafischen Elemente in Diskussionsbeiträgen nicht abdrucken.
- Füllen Sie eine Erklärung zu einem möglichen Interessenkonflikt aus.
- Bearbeiten Sie die deutschen und englischen Satzfahnen nach Erhalt ohne Verzögerung.
- Geben Sie eine Adresse an. Anonyme Diskussionsbeiträge können wir nicht publizieren.
- Senden Sie Ihren Diskussionsbeitrag zu Artikeln der Medizinisch-Wissenschaftlichen Redaktion an: medwiss@aerzteblatt.de oder Deutsches Ärzteblatt, Dieselstraße 2, 50859 Köln.

Zusatzmaterial zu:

Qualität und Nutzen künstlicher Intelligenz in der Patientenversorgung

Kai Wehkamp, Michael Krawczak, Stefan Schreiber

Dtsch Arztebl Int 2023; 120: 463–9. DOI: 10.3238/arztebl.m2023.0124

eLiteratur

e1. Haenssle HA, Fink C, Toberer F, et al.: Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020; 31: 137–43.

e2. Blaha J, Barteczko-Grajek B, Berezowicz P, et al.: Space Glucose-Control system for blood glucose control in intensive care patients—a European multicentre observational study. BMC Anesthesiol 2016; 16: 8.

e3. Repici A, Badalamenti M, Maselli R, et al.: Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020; 159: 512–20.e7.

e4. Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M: Stand-alone use of artificial intelligence for digital mammography and digital breast tomosynthesis screening: a retrospective evaluation. Radiology 2022; 302: 535–42.

e5. Meyer A, Zverinski D, Pfahringer B, et al.: Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018; 6: 905–14.

e6. Braun T, Spiliopoulos S, Veltman C, et al.: Detection of myocardial ischemia due to clinically asymptomatic coronary artery stenosis at rest using supervised artificial intelligence-enabled vectorcardiography—a five-fold cross validation of accuracy. J Electrocardiol 2020; 59: 100–5.

e7. Ipp E, Liljenquist D, Bode B, et al.: Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open 2021; 4: e2134254.

e8. Franzke AW, Kristoffersen MB, Bongers RM, et al.: Users' and therapists' perceptions of myoelectric multi-function upper limb prostheses with conventional and pattern recognition control. PLoS One 2019; 14: e0220899.

e9. Daifalla K, Günther S: Eigener Report 2022: Mindpeak Breast HER2 RoI Clinical Performance Evaluation Summary. www.uploads-ssl.webflow.com/60424989e8e0f02a922616f9/631072d2e19725a967c1735f_Mindpeak%20Breast%20HER2%20RoI%20-%20Clinical%20performance%20evaluation%20summary%20-%20APPROVED.pdf (last accessed on 20 November 2022).

e10. Sun H, Depraetere K, Meesseman L, et al.: Machine learning-based prediction models for different clinical risks in different hospitals: evaluation of live performance. J Med Internet Res 2022; 24: e34295.

e11. Chamberlin J, Kocher MR, Waltz J, et al.: Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value. BMC Med 2021; 19: 55.

eKASTEN

Erläuterungen ausgewählter Begriffe und Konstrukte

Big Data
Der Begriff Big Data bezeichnet große und oft unstrukturierte Datenmengen, die aufgrund ihres Umfangs und ihrer Komplexität nicht mehr ohne Weiteres durch Menschen oder einfache Algorithmen verarbeitet und interpretiert werden können. Anwendungen des → maschinellen Lernens werden dafür entwickelt, um mit großen Datensätzen trainiert zu werden und hierin Muster zu erkennen. Es besteht die Hoffnung, dass diese Anwendungen künftig bislang unerkannte Muster in komplexen medizinischen Daten erkennen (zum Beispiel bestimmte medizinische Zusammenhänge und Risikofaktoren). Ein Beispiel für große Datenmengen bietet der Bereich → Multiomics.

„Deep learning", „deep neuronal network" (DNN)
Mit „deep learning" wird ein bestimmtes Verfahren des → maschinellen Lernens bezeichnet, das auf sogenannten tiefen neuronalen Netzen beruht. Kennzeichnend ist eine äußere Eingabeschicht, die die zu verarbeitenden Informationen erfasst (zum Beispiel die Pixel eines Röntgenbildes). Die Einzelinformationen der Eingabe werden über digitale Verknüpfungen (als Neurone bezeichnet) über mehrere Ebenen von Knotenpunkten gewichtet, verknüpft und weitergeleitet, um letztlich als Ausgabe eine Verarbeitung beziehungsweise Interpretation der ursprünglichen Eingabeinformation zu erzielen (zum Beispiel Klassifikation eines Rundherdes im Röntgenbild). Von einem tiefen neuronalen Netzwerk spricht man aufgrund der Vielzahl von Zwischenebenen. Beim überwachten Lernen entspricht das Training eines neuronalen Netzes vereinfacht dargestellt einer Gewichtung der digitalen Neurone. Diese wird dadurch erzielt, dass auf Eingabeebene eine Vielzahl an Beispielen präsentiert wird (zum Beispiel Röntgenbilder mit und ohne Rundherd), die auf die vorgegebene Ausgabe (zum Beispiel „dies ist ein Rundherd", „dies ist ein Normalbefund") optimiert werden. Die Verknüpfungen und Gewichtungen der Neurone werden im Rahmen des Trainings durch das System immer weiter angepasst und haben eine Komplexität und Tiefe, die in der Regel nicht mehr transparent gemacht werden kann. Das trainierte neuronale Netzwerk ist dann optimalerweise dazu in der Lage, bisher unbekannten Eingaben (zum Beispiel neue Röntgenbilder) mit der richtigen Ausgabe zu verknüpfen.

Konformitätsbewertung für Medizinprodukte
Medizinprodukte, zu denen auch auf maschinellem Lernen basierende Anwendungen gehören, müssen ihre Konformität mit dem deutschen Medizinproduktegesetz nachweisen (das wiederum die Regelungen der europäischen Medizinprodukteverordnung [MDR] umsetzt). Je nach → Risikoklasse des Medizinprodukts muss der Hersteller unter anderem eine technische Dokumentation und ein spezifisches Qualitätsmanagementsystem vorlegen. Inhaltlich muss gezeigt werden, dass die dargestellten Funktionen durch das Produkt erfüllt werden. Bislang müssen in Europa für die Konformitätsbewertung der gängigen KI-Anwendungen keine klinischen Anwendungsstudien vorgelegt werden. Es wird aber diskutiert, zukünftig – ähnlich wie in den USA – klinische Studien für die Zulassung bestimmter KI-Anwendungen zu fordern.

Maschinelles Lernen
Maschinelles Lernen (ML) bezeichnet eine bestimmte Technik, um künstliche Intelligenz zu erzeugen. Kennzeichnend ist dabei ein digitales, technisches Lernen von Mustern anhand von Datensätzen, mit dem Ziel, diese Muster zur Interpretation neuer, bislang unbekannter Daten anzuwenden. Überwachtes Lernen, unüberwachtes Lernen und bestärkendes Lernen sind Untergruppen des maschinellen Lernens *(Grafik 1)*. Es gibt verschiedene Techniken, um Systeme des maschinellen Lernens zu programmieren. Hierzu gehören beispielsweise die tiefen neuronalen Netze (→ „deep learning") oder auch die sogenannten Entscheidungswälder.

Multiomics
Multiomics bezieht sich auf die Zusammenführung und Interpretation verschiedener biologischer Kategorien von Daten, das heißt dem Genom, Transkriptom, Proteom, Metabolom, Epigenom, Mikrobiom und weiteren. Der Begriff leitet sich von den gemeinsamen letzten Silben der englischsprachigen Begriffe der technologischen Bereiche ab (zum Beispiel „Genomics"). Aktuell entwickelt sich in diesem Zusammenhang ein bedeutender Forschungszweig. Es besteht die Hoffnung durch eine Verknüpfung dieser großen Datenmengen (→ Big Data), in Verbindung mit maschinellem Lernen bislang unbekannte Mechanismen und Assoziationen für Risikofaktoren und Erkrankungen zu identifizieren.

„Natural language processing"
„Natural languagep processing" (NLP) bezeichnet digitale Techniken zur Verarbeitung und Interpretation von Sprache. Heutzutage basiert NLP in der Regel auf → „deep learning", das heißt tiefen neuronalen Netzen als Technik des → maschinellen Lernens (ML). Die Systeme werden dabei mit Text- oder Audiodaten trainiert. In komplexen ML-Anwendungen werden Textdaten (zum Beispiel medizinische Briefe) teilweise mittels NLP vorverarbeitet, um dann in einem weiteren Schritt mit anderen Daten zusammengeführt zu werden.

Risikoklassen Medizinprodukte

Entsprechend der europäischen und deutschen Rechtsverordnungen werden vier verschiedene Risikoklassen für Medizinprodukte unterschieden, in die auch Anwendungen des maschinellen Lernens eingeordnet werden:

- Klasse I: niedriges Risiko (zum Beispiel Lesebrillen, Software zur Zyklusbestimmung)
- Klasse IIa: mittleres Risiko (zum Beispiel Ultraschallgeräte, Software zur medizinischen Diagnoseunterstützung ohne unmittelbare Gefahr für die Patienten)
- Klasse IIb: hohes Risiko (zum Beispiel Beatmungsgeräte, Software, die vitale Funktionen kontrolliert) und
- Klasse III: sehr hohes Risiko (zum Beispiel Arzneimittel, autonome Softwaresysteme die bei Fehlfunktion direkt zum Tod führen können).

Je nach Stufe werden unterschiedliche Anforderungen für die → Konformitätsbewertung vorausgesetzt. Viele auf maschinellem Lernen basierende Anwendungen dienen formal nur der Unterstützung ärztlicher Entscheidungen, das heißt die Verantwortung bleibt beim Menschen. Dementsprechend müssen hier nur die weniger anspruchsvollen Sicherheitskriterien nach Klasse I oder IIa erfüllt werden müssen, sodass sich Risiken ergeben, wenn diese Systeme in der praktischen Anwendung doch ohne menschliche Kontrolle eingesetzt werden.

Fragen zu dem Beitrag aus Heft 27–28/2023:

Qualität und Nutzen künstlicher Intelligenz in der Patientenversorgung

cme plus

Einsendeschluss ist der 09.07.2024. Pro Frage ist nur eine Antwort möglich.
Bitte entscheiden Sie sich für die am ehesten zutreffende Antwort.

Frage Nr. 1

Welche Aussage zu den Beziehungen zwischen menschlicher Intelligenz, künstlicher Intelligenz und „machine learning“ ist richtig?

a) Künstliche Intelligenz ist heute allen Facetten menschlicher Intelligenz überlegen.
b) Durch die Entwicklung tiefer neuronaler Netze („deep learning“) wurden im Bereich „machine learning“ in den letzten Jahren große Fortschritte erreicht.
c) Ein Schachcomputer verfügt in der Regel nicht über künstliche Intelligenz.
d) Allgemein wird erwartet, dass künstliche Intelligenz bis zum Ende des aktuellen Jahrzehnts die menschliche Intelligenz in allen Bereichen übertreffen wird.
e) Die typischerweise von „machine learning“-Anwendungen bearbeiteten Aufgaben, können in der Regel schneller und genauer durch Menschen bearbeitet werden.

Frage Nr. 2

Für eine „machine learning“-Anwendung werden in einem großen Datensatz von Mammografie-Bildern jeweils Malignom-suspekte Befunde durch einen erfahrenen Radiologen gekennzeichnet. Anhand dieser Bilder wird ein Algorithmus trainiert und das hieraus entstandene Muster an neuen, bislang nicht bearbeiteten Mammografie-Bildern überprüft. Es zeigt sich, dass auffällige Befunde mit einer hohen Präzision erkannt werden. Um welche Art des „machine learning“ handelt es sich hier?

a) unüberwachtes Lernen
b) bestärkendes Lernen
c) genetisches Lernen
d) überwachtes Lernen
e) erklärendes Lernen

Frage Nr. 3

Ein „machine learning“-Algorithmus soll darauf trainiert werden, epidemische Infektionsausbrüche in den an die Gesundheitsämter gemeldeten infektiologischen Daten zu erkennen. Hierfür muss der Algorithmus lernen, zwischen dem Grundrauschen sporadisch vorkommender Infektionen und einem sich manifestierenden Ausbruchsgeschehen zu differenzieren.
Um welche Art des „machine learning“ handelt es sich hier?

a) unüberwachtes Lernen
b) bestärkendes Lernen
c) genetisches Lernen
d) überwachtes Lernen
e) erklärendes Lernen

Frage Nr. 4

Der „machine learning“-Zyklus stellt die Abhängigkeit der verschiedenen voneinander abhängigen Ebenen dar. Welche Abfolge gibt Inhalt und Reihenfolge passend wieder?

a) Programmierung → Sammeln von Daten → Testung → Erstellung einer Datenbank → Programmierung
b) reale Lebenswelt → Erstellung von Zufallsdaten → Programmierung → Erstellung von Anwendungsdaten
c) Recherche, Diagnose-Score → Programmierung, Datenerkennung → Kalkulation angewandter Diagnose-Scores → Publikation
d) Anwendung → Rückkopplung → Datenvalidierung → Testumgebung → Anwendung
e) reale Lebenswelt → digitale Daten → Vorbereitung von Variablen → Design des KI-Algorithmus, Anwendung → reale Lebenswelt

Frage Nr. 5

Welches Argument wird im Text aufgeführt, um zu begründen, dass Medizinerinnen und Mediziner sich mit den Konzepten künstlicher Intelligenz auseinandersetzen sollten?

a) Risiken von KI-Anwendungen beurteilen zu können
b) Eindruck bei Patientinnen und Patienten sowie Angehörigen zu machen
c) eigene KI-Anwendungen entwickeln zu können
d) Schnittstellen programmieren zu können
e) Daten zu sammeln

Frage Nr. 6

Wie viele medizinische KI-Anwendungen sind in den USA in etwa bereits durch die FDA zugelassen?

a) ca. 10
b) ca. 55
c) ca. 180
d) ca. 520
e) ca. 1 200

Frage Nr. 7

In einem Krankenhaus werden verschiedene Systeme künstlicher Intelligenz im Bereich der Diagnostik eingesetzt. Wer trägt in der Regel die Verantwortung für die darauf basierenden Behandlungsentscheidungen?

a) der Hersteller der KI-Anwendung
b) die kaufmännische Geschäftsführung
c) die behandelnden Ärztinnen und Ärzte
d) die künstliche Intelligenz
e) das Gewerbeaufsichtsamt

Frage Nr. 8

Welche Art des KI-Lernens funktioniert nach dem Verfahren „trial and error“?

a) überwachtes Lernen
b) unüberwachtes Lernen
c) wachsendes Lernen
d) unselbständiges Lernen
e) bestärkendes Lernen

Frage Nr. 9

Welche Art von Validierung für den Nutzen einer neuen KI-basierten medizinischen Anwendung wird im Text gefordert?

a) prospektive Interventionsstudien
b) retrospektive Datenbankanalysen
c) Fragebogen zur Zufriedenheit der Ärztin/des Arztes
d) Studien in verschiedenen Ländern
e) Querschnittstudien

Frage Nr. 10

Bei welchem der folgenden Beispiele im Beitrag war die Detektionsrate in der Diagnostik mit KI höher als ohne KI?

a) Präeklampsie
b) kolorektale Neoplasie
c) feuchte Makuladegeneration
d) Sjögren-Syndrom
e) Aufmerksamkeits-Defizit/Hyperaktivitäts-Störung (ADHS)