Maschinelles Lernen in der Anästhesiologie – Anwendungen, Entwicklungsprozess und Ausblick

Machine learning in anaesthesiology – existing models and steps towards clinical application

A. Luckscheiter[1] · W. Zink[1] · M. Thiel[2] · V. Schneider-Lindner[2]

▶ **Zitierweise:** Luckscheiter A, Zink W, Thiel M, Schneider-Lindner V: Maschinelles Lernen in der Anästhesiologie – Anwendungen, Entwicklungsprozess und Ausblick.
Anästh Intensivmed 2024;65:466–478. DOI: 10.19224/ai2024.466

Zusammenfassung

Hintergrund

Die Generierung großer elektronischer Datenmengen der Anästhesiologie (Anästhesie, Intensivmedizin, Notfallmedizin, Schmerz- und Palliativmedizin) ermöglicht die Entwicklung von Modellen des maschinellen Lernens (ML). Diese haben das Potenzial, einen integralen Bestandteil der zukünftigen Patientenversorgung zu bilden.

Ziel der Arbeit

Die Übersichtsarbeit soll mögliche Anwendungsfelder von ML-Modellen aufzeigen und so deren potenziellen Einfluss auf die klinische Arbeit in der Anästhesiologie verdeutlichen. Anhand fiktiver Fälle sollen aktuelle Studien die Chancen und Risiken des ML in der Anästhesiologie sowie die Herausforderungen hin zur finalen klinischen Anwendung aufzeigen.

Material und Methoden

Es erfolgte eine selektive PubMed-Recherche zu den Stichworten Anästhesiologie, Intensivmedizin, Notfallmedizin sowie Schmerz- und Palliativtherapie zusammen mit maschinellem Lernen. Die Studienauswahl erfolgte anhand einer möglichen baldigen oder tatsächlichen klinischen Anwendung, zur Schaffung eines Problembewusstseins oder anhand eines vermehrten klinischen Bedarfs an ML-Unterstützung bezogen auf Hauptschwerpunkte der alltäglichen Versorgung in den jeweiligen Fachbereichen. Sofern verfügbar, wurden diese mit etablierten Scores verglichen.

Ergebnisse

In der Anästhesie konnten ML-Modelle zur Vorhersage des schwierigen Atemwegs, Hypoxie, Hypotonie und Narkosetiefe identifiziert werden. In der Intensivmedizin könnten Modelle durch Vorhersagen zu Mortalität, Morbidität und Organdysfunktion unterstützend wirken. Limitierte Datenquellen schränken das ML in der präklinischen Notfallmedizin ein. Allerdings sind hier bereits Modelle zum Reanimationserfolg, zum Intubationsrisiko sowie zur schnelleren EKG-Diagnostik erschienen. In der Schmerz- und Palliativmedizin könnte die Diagnostik und Therapie von neuropathischen Schmerzzuständen verbessert bzw. die Ressourcensteuerung bei palliativer Versorgung optimiert werden.

Diskussion

In der Anästhesiologie könnten viele Aspekte der Patientenversorgung durch ML-Anwendungen unterstützt werden. Allerdings umfasst die Entwicklung von ML-Modellen nicht nur die Verifikation der Datenvalidität, Schulungen oder interne und externe Testungen. Weitere Schritte bis zur Implementierung betreffen Validierungsstudien, die Einhaltung ethischer, rechtlicher und technischer Standards sowie Anwenderschulungen und Überwachungsprozesse. Daher hat maschinelles Lernen in der Anästhesiologie noch nicht sein volles Potenzial erreicht.

1 Klinik für Anästhesiologie, Operative Intensivmedizin und Notfallmedizin, Klinikum Ludwigshafen (Direktor: Prof. Dr. W. Zink)

2 Klinik für Anästhesie, Operative Intensivmedizin und Schmerzmedizin, Universitätsklinikum Mannheim (Direktorin: Prof. Dr. G. Beck)

Anmerkung

Im Manuskript wird aus Gründen der Lesbarkeit auf das Gendern verzichtet und das grammatikalisch korrekte Genus verwendet. Sofern nicht explizit vermerkt, sind immer alle Geschlechter gleichermaßen damit gemeint.

Interessenkonflikt

Die Autorinnen und Autoren geben an, dass keine Interessenkonflikte bestehen.

Schlüsselwörter
Notfallmedizin – Intensivmedizin – Anästhesiologie – Schmerztherapie – Palliativmedizin – Maschinelles Lernen

Keywords
Anaesthesiology – Intensive Care Medicine – Emergency Medicine – Pain Therapy – Palliative Care – Machine Learning

Summary

Background

Big data from anaesthesiology, (anaesthesia, emergency medicine, intensive care medicine, pain and palliative therapy) increasingly facilitates the development of machine learning-based tools and models which have a strong potential to improve patient care.

Objectives

The aim of the study is to highlight application areas for machine learning (ML) and to describe the potential impact of machine ML-based tools on clinical work in anaesthesiology. Based on fictitious clinical cases a selection of current, applied ML studies will be presented as well as the challenges that need to be met prior to the clinical application of ML models.

Material and methods

A PubMed query including the keywords anaesthesiology, intensive care medicine, emergency medicine, pain therapy, palliative care and machine learning was performed. With regard to key topics of the specialties, we selected studies describing ML analyses likely to lead to future clinical applications or meet the growing clinical requirement of an ML-based support. Where available, the study results were compared to established clinical scores.

Results

In anaesthesia, ML-based tools could support the prediction of a difficult airway, hypotension, hypoxia or depth of anaesthesia. In intensive care medicine, mainly prediction models for mortality, morbidity and organ failure exist. The role of ML in emergency medicine is still limited due to restricted data resources. However, the success of resuscitation and the need of airway management can be modelled; ECG diagnosis can be improved. Pain therapy can be optimised and neuropathic pain can be predicted. A better individualised resource management in palliative care could be achieved by means of prognostic forecasting with ML.

Conclusion

In anaesthesiology many aspects of clinical care could be supported with ML-based tools. However, subsequent to the underlying ML model development, which comprises verification of data validity, training, internal and external testing, implementation of ML-based tools requires the completion of significant further steps including validation studies, compliance with ethical, legal and technical standards, user training and governance. Therefore, exemplified by ML-based clinical tools, its full potential in anaesthesiology has not yet been reached by far.

Hinführung

Der folgende Übersichtsartikel soll einen Überblick über Einsatzmöglichkeiten des maschinellen Lernens aus großen Datenmengen in der Anästhesie, Inten-

siv- und Notfallmedizin sowie Schmerz- und Palliativtherapie vermitteln und notwendige Entwicklungsschritte für zukünftige Anwendungen aufzeigen. Mittels fiktiver Fälle sollen ein praxisorientierter Ansatz aufgezeigt und mögliche Anwendungs- und Entwicklungsprobleme aufgezeigt werden.

Einleitung

Obwohl bereits seit der letzten Dekade maschinelles Lernen (ML) und künstliche Intelligenz (KI) viele Lebensbereiche und das Gesundheitswesen beeinflussen, ist erst seit der Verbreitung von Programmen wie Chatbots dieses Thema vielen Menschen im Alltag präsent [1,2]. Dabei stellt ML nur einen Teilbereich der KI dar, bei dem maschinell lesbare Daten von speziellen, nicht spezifisch programmierten Algorithmen ausgewertet werden. Aus Trainingsdaten resultiert so ein statistisches Modell, das Muster und Gesetzmäßigkeiten für Vorhersagen oder Eingruppierungen (Clustering) erkennt. Allgemein unterscheidet man ein überwachtes Lernen, bei dem die Zielgröße im Trainingsverfahren vorgegeben wird, von einem unüberwachten Lernen ohne bekannte Zielwerte, um Datenmengen in Untergruppen zu unterteilen. KI wiederum nutzt ML-Algorithmen zur Programmgenerierung, um durch Einspeisung von Daten aus seiner Umgebung, deren Speicherung und Interpretation mit seiner Umwelt vorrausschauend und angemessen zweckgebunden zu interagieren [3,4]. Aufgrund möglicher tiefgreifender Alltagsveränderungen ist eine breite Diskussion über Chancen und Risiken dieser Anwendungen entstanden, welche oftmals stark emotional geführt wird [5]. In einer aktuellen Umfrage in der Anästhesiologie zeigte sich die Mehrheit skeptisch gegenüber der Digitalisierung im Gesundheitswesen [3,6]. Gründe hierfür waren u. a. hohe Kosten bei fehlender Refinanzierung, mangelnde Qualität der Software und des Service sowie unzureichende Praktikabilität [7]. Bei jedoch neutraler Überprüfung fällt auf, dass KI in der Medizin sich erst am Anfang befindet und der Weg bis zum breiten klinischen Einsatz noch weit ist. Daher möchte dieser Artikel zunächst eine praktische Übersicht über exemplarische ML-/KI-Anwendungen zu Kernthemen der Anästhesie, Intensiv- und Notfallmedizin sowie Schmerz- und Palliativtherapie (AINSP) geben. Gleichzeitig soll hieran der Weg zum klinisch zugelassenen Programm aufgezeigt werden. Damit soll der Leser eine sachliche Vorstellung über den potenziellen Einfluss maschinellen Lernens im zukünftigen Arbeitsalltag erhalten [8].

Um die Chancen und Risiken des maschinellen Lernens besser zu begreifen und um Studien hierzu objektiver interpretieren zu können, sollen zunächst einige allgemeine Vorbetrachtungen ausgeführt werden. Für weitere Grundlagen des maschinellen Lernens sowie bezüglich der einzelnen Algorithmen soll an dieser Stelle auf weitere Literatur verwiesen werden [3,4,9]. Alle ML-Algorithmen berechnen Wahrscheinlichkeiten, die sie mit einem vorher festgelegten Schwellenwert (Threshold) abgleichen. Wird er unterschritten, wird dagegen votiert, wird er überschritten, dafür. Durch die Anpassung des Schwellenwerts können unterschiedlich hohe Sensitivitäten/Spezifitäten entstehen. Dagegen darf der positive/negative prädiktive Wert (PPV/NPV) wegen der Prävalenzabhängigkeit nur unter Betrachtung der Repräsentanz und der Kohortengröße interpretiert werden. Daher müssen weitere Testgütekriterien wie die Fläche unter der Grenzwertoptimierungskurve (AUC-ROC, AUC – **area under the curve,** ROC – **receiver operating characteristic**) sowie die Precision-Recall-Kurve (AUC-PRC) zur Ergebnisinterpretation genutzt werden (Tab. 1). Bei der AUC-ROC werden zur graphischen Trennschärfe in Abhängigkeit verschiedener Schwellenwerte die Sensitivität auf der Ordinate gegenüber der falsch positiven Rate (Abszisse) aufgetragen. Die vom Nullpunkt zum Punkt 1/1 hin gekrümmte Kurve wird wie folgt interpretiert: Ist sie gleich einer Diagonalen vom Nullpunkt zum Punkt 1/1, so liegt eine zufällige Klassifikation vor (AUC-ROC = 0,5), je mehr sie sich gegen 1/0 (und damit zur linken oberen Ecke) hin krümmt, desto höher ist die Trennschärfe (Tab. 1) [4,10]. Nachteile bei der AUC-ROC können sich dann ergeben, wenn die Grundwahrscheinlichkeit eines Zustandes in einer Kohorte unverhältnismäßig höher repräsentiert ist und Spezifität und Sensitivität zudem relativ hoch sind (z. B. Verhältnis intubierte zu nicht intubierten Polytraumapatienten 1:24, Sensitivität 0,73, Spezifität 0,99, AUC-ROC 0,96 [11]). Hier kann die AUC-ROC eine extrem gute Testgüte vortäuschen, ohne dabei die Prävalenz und damit den PPV zu berücksichtigen [12,13]. Die Precision-Recall-Kurve vergleicht PPV (auf Englisch auch als **precision** bezeichnet, Ordinate) zur Sensitivität (Englisch **recall,** Abszisse) unter Berücksichtigung verschiedener Schwellenwerte. Sie zieht gekrümmt vom linken oberen Quadranten Punkt 0/1 zum Punkt 1/0. Eine AUC von nahe 1,0, d. h. eine Kurve, die zur rechten oberen Ecke (Punkt 1/1) hin tendiert, zeigt somit eine hohe Wahrscheinlichkeit für ein Kombination aus hoher Sensitivität und PPV. Bei der AUC-PRC ist eine zufällige Vorhersage jedoch abhängig von der Prävalenz, besser gesagt von der Klassenbalance. In einem ausgeglichenen Datenset (Ratio 1:1) entspräche eine zufällige Vorhersage einer AUC von 0,5 (gedacht als parallele Gerade zur x-Achse), bei einer Imbalance von bspw. 1:10 einer AUC von nur 0,09 [13]. AUC-ROC und AUC-PRC sollten gemeinsam betrachtet und interpretiert werden. Eine ROC-Analyse ist bspw. zielführender, wenn die Sensitivität höher eingestuft wird (z. B. Erfassung aller Spam-E-Mails) wohingegen eine PRC-Analyse vor allem bei Betrachtung der Prävalenz bzw. PPV wichtiger erscheint (z. B. Test auf eine seltene Infektionskrankheit). Einen Hinweis auf die Ratio zwischen zwei Klassen (Klassenimbalance) können Maße wie der F-Score oder Matthews' Korrelationskoeffizient geben [14].

Auch Datenakquise und Attributsauswahl sind für die Aussagekraft einer ML-Analyse essenziell. Für alle verwendeten Attribute (z. B. **Glasgow Coma Scale** (GCS) mit Werten 3–15) muss sichergestellt sein, dass sie mit höchster Sorgfalt bezüglich medizinischer Plau-

Tabelle 1
Auswahl wichtiger Gütekriterien von Modellen des Maschinellen Lernens [4,10,12–14].

Gütekriterium	Wert	Beschreibung	Bemerkungen
Sensitivität (Recall)	0,0–1,0	Anteil der korrekt positiv Getesteten an allen tatsächlich Positiven	
Positiv prädiktiver Wert (PPV)	0,0–1,0	Anteil der korrekt positiv Getesteten an allen positiv Getesteten	Prävalenzabhängig
Spezifität	0,0–1,0	Anteil der korrekt negativ Getesteten an allen tatsächlich Negativen	
Negativ prädiktiver Wert (NPV)	0,0–1,0	Anteil der korrekt negativ Getesteten an allen negativ Getesteten	Prävalenzabhängig
Grenzwertoptimierungskurve (area under the receiver-operating characteristic-curve, AUC-ROC)	0,5–0,6 unbefriedigend 0,6–0,7 befriedigend 0,7–0,8 gut 0,8–0,9 sehr gut >0,9 exzellent nach [10]	Graphische Darstellung der Diskriminierungs-wahrscheinlichkeit eines Tests, in dem Sensitivität und falsch positive Rate (= 1-Spezifität) bei Verwendung verschiedener Schwellenwerte aufgetragen werden	• Je näher an 1, desto genauer das Modell • Wird durch die Ratio von positiv und negativ in der Kohorte stark beeinflusst, da bei gleichzeitiger hoher Spezifität eine zu hohe Testdiskriminierung vorgetäuscht wird • Theoretisch Werte <0,5 möglich
Precision-Recall-Kurve **(area under the precision-recall-curve, AUC-PRC)**	0,0–1,0	Graphische Darstellung des Verhältnisses von Sensitivität zu PPV bei Verwendung verschiedener Schwellenwerte	• Je näher an 1, desto genauer das Modell • Kann auch für Spezifität und NPV erstellt werden • Prävalenzabhängiger Wert für eine Zufallsaussage
F-Maß	Umfang 0 = keine Testgüte, 1 = perfekte Sensitivität und PPV	Verhältnis von Genauigkeit (PPV) und Trefferquote (Sensitivität) als gewichtetes harmonisches Mittel	• Hinweise auf Klassenimbalance • Kann auch für Spezifität und NPV erstellt werden
Matthews-Korrelations-koeffizient	Umfang -1 = keine Übereinstimmung zwischen Vorhersage und Beobachtung, 0 = Zufallsvorhersage, +1 = totale Übereinstimmung	Misst die Differenz zwischen den vorhergesagten und den tatsächlichen Werten ähnlich dem Chi-Quadrat-Test	• Besser geeignet bei binären Vorhersagemodellen [6] • Umfasst immer das Modell als ganzes
Gesamtkorrektheit	Angabe in Prozent		• Kann durch eine Klassenimbalance eine zu hohe Testgenauigkeit vortäuschen

sibilität ausgewählt und ihre Messwerte reproduzierbar und valide erfasst wurden. Andernfalls hat das Modell bei einer externen Validierung eine schlechte oder gar keine Aussagekraft. Zudem dürfen Daten zu Geschlecht, Ethnie oder Religion nur dann zur Berechnung herangezogen werden, wenn es eine wissenschaftlich fundierte, pathophysiologische oder medizin-soziologische Begründung hierfür gibt. Ansonsten erhöht sich die Gefahr der (unbewussten) Diskriminierung, gerade bei Unterrepräsentanz einzelner Subgruppen. Standardisiert erhobene oder objektiv bestimmbare Attribute sind die Grundlage für besser reproduzierbare Ergebnisse. Unter Betrachtung dieser Prämisse ergibt sich für die Validität eines Modells, dass Attribute aus physikalisch-biochemischen Messungen eine bessere Reproduzierbarkeit aufweisen können als klinisch-subjektive Merkmale oder patientenindividuelle Sichtweisen. Dies erschwert KI-Anwendungen in Themengebieten, die ethische Wertvorstellungen berücksichtigen müssen [15,16]. Dagegen sind in der Anästhesie und Intensivmedizin durch die digitale Erhebung und Prozessierung der Vitalparameter, radiologischen Befunde und Laborwerte die Voraussetzungen für die Nutzung von ML-Methoden für Forschung und Krankenversorgung günstig. Notfallmedizin und Schmerztherapie haben hier eine Mittelstellung inne, z. T. wegen mangelnder Digitalisierung und klinisch subjektiver Befunde. Grundsätzlich kann ML für verschiedene Zielsetzungen verwendet werden: frühzeitigere Detektion von Krankheiten (z. B. Differenzialdiagnosen), Zustandsverschlechterungen (z. B. Organversagen), Phänotypisierungen (z. B. Sepsis-Subtypen), individualisierte Entscheidungsprozesse (z. B. Beatmungsstrategien) und Vergleich mit Falldatenbanken zur individuellen Versorgung [17].

Material und Methoden

Es wurde eine selektive PubMed-Recherche nach Studien zu den Kernthemen maschinelles Lernen sowie Anästhesie, Intensiv- und Notfallmedizin sowie Schmerz- und Palliativtherapie durchgeführt (Schlagworte: Akutes Lungenversagen, Reanimation, Awareness, Organversagen, Polytrauma, Sepsis, Atemwegsmanagement, Neuralgie, Neuropathischer Schmerz, Herpes Zoster; für Details s. Kapitel „Literaturrecherche“ in der Online-Ausgabe des Artikels). Eine

Filterung respektive ein Mindestzeitraum seit der Erstpublikation wurde nicht gewählt. Dieser Artikel möchte seinen Fokus auf eine einfache Lesbarkeit legen, um ein breites Publikum zu erreichen. Daher erhebt er keinen Anspruch auf vollständige Abbildung der Studienlage zu den einzelnen Suchbegriffen. Vielmehr wurde als Auswahlkriterium nicht nur eine möglichst hohe klinische Relevanz der exemplarischen Studien in Kernthemen der AINSP verwendet. Zudem sollten mit beispielhaften Untersuchungen der Entwicklungsstand, Probleme in der Entstehung und Validierung sowie Ausblicke auf künftig mögliche Entwicklungen dargestellt werden. Darüber hinaus wurden Studien zu den kommerziell nutzbaren und von der Food and Drug Administration (FDA) zugelassenen Modellen selektiert. Die Testgütekriterien für zum Vergleich herangezogene nicht ML- bzw. KI-basierte Scoring-Systeme wurden entweder händisch gesucht oder aus den exemplarischen Studien entnommen. Für weitere Informationen sei auf die themenspezifisch aufgeführten Übersichtsartikel verwiesen, aus denen ebenfalls einzelne Studien selektiert wurden.

Ergebnisse

Fall 1 – Anästhesie

Ein 65-jähriger Patient stellt sich zur elektiven Implantation einer Hüftendoprothese vor. An Vorerkrankungen weist der adipöse Patient einen arteriellen Hypertonus, einen oral eingestellten Diabetes mellitus sowie eine beginnende chronisch obstruktive Lungenkrankheit bei Nikotinabusus auf. Im anästhesiologischen Aufklärungsgespräch entscheidet sich der Patient für eine Allgemeinanästhesie (Mallampati 3, eingeschränkte Reklination). Inwieweit könnte maschinelles Lernen helfen, Patientensicherheit in der präoperativen Evaluation sowie während der Narkose zu erhöhen?

Die präoperative Evaluation der patientenindividuellen Risiken dient zur Vermeidung von Komplikationen während der Anästhesie. Ein wichtiger Aspekt ist die Beurteilung des Atemwegs. Allerdings liegt die Sensitivität des besten klinischen Tests für eine schwierige Intubation (modifizierter Mallampati-Test) nur bei 0,51 [18]. Tavolara et al. konnten durch ein neuronales Netzwerk, welches Frontalaufnahmen der Patientengesichter analysierte, eine deutlich höhere Klassendiskriminierung erzielen als klassische Scores (AUC-ROC 0,71 vs. 0,47 für thyromentaler Abstand ≤3 Fingerbreiten bzw. 0,6 für Mallampti-Score ≥3) [19]. Die Sensitivität reichte von 0,90 (Spezifität 0,44) bis zu 0,36 (Spezifität 0,96). Eine Erweiterung um Seitaufnahmen, so die Autoren, könnte zu einer weiteren Verbesserung beitragen. Künftig könnte die Patientensicherheit durch die Vorhersage eines unerwartet schwierigen Atemwegs somit über ein einfaches Patientenfoto erhöht werden. Da jedoch die Studiengröße mit zwei Gruppen zu je n = 76 zu gering und nicht ausbalanciert war, müsste vor einem weiteren klinischen Einsatz erst die Datenbasis vergrößert werden.

Die Vermeidung intraoperativer Hypotonie- und Hypoxiephasen trägt wesentlich zur Aufrechterhaltung der intraoperativen Homöostase und höchstwahrscheinlich auch zu verminderten postoperativen Komplikationen bei. So gehen intraoperative Hypotonien mit dem vermehrten Auftreten von postoperativen Schlaganfällen einher [20]. Wijnberge et al. konnten mit Hilfe eines auf Blutdruckmessung und Pulskonturanalyse basierenden Modells zeigen, dass die intraoperative Hypotoniezeit (mittlerer arterieller Druck (MAP) <65 mmHg für eine Minute) signifikant von 32 min auf 8 min bei nicht kardiochirurgischen Eingriffen reduziert werden konnte [21]. Nach externer Validierung ging daraus der industriell vermarktete **Hypotension Prevention Index** hervor (HPI) (Edwards Lifesciences Corporation, Irvine, Kalifornien, USA, Sensitivität 0,88–0,92, Spezifität 0,87–0,92, AUC-ROC 0,95–0,97), dessen Evidenz für harte Endpunkte wie kardioembolische Ereignisse jedoch limitiert ist [22,23]. Allerdings existieren Zweifel an der hohen Spezifität. Vermutet wird, dass durch die Deklaration der Normotension (MAP >75 mmHg für 30 min mit mindestens 20 min Abstand von jeder hypotensiven Episode) ein Selektionsbias entstanden ist und der HPI generell einen MAP <75mmHg als zukünftige Hypotension vorhersagt [24]. Zur Vermeidung intraoperativer Hypoxien stellten Lundberg et al. [90] ein Modell vor, welches während der Narkose zuverlässiger Entsättigungen vorhersagte als die reinklinische Erfahrung. Sie nutzten Attribute von mehr als 50.000 Patienten wie Nüchternheit, Body-Mass-Index, initiale Sättigung, aber auch Werte des Beatmungsgeräts wie positiv endexspiratorischer Druck, Tidalvolumen, endtidales CO_2, Kreislaufwerte und Sauerstoffsättigung, und berechneten für ein Fünfminutenintervall das Risiko für einen Sättigungsabfall <93 %. Für eine Hypoxie nach Einleitung lag die AUC-ROC bei 0,6 (Anästhesist), 0,76 (KI und Anästhesist) bzw. 0,83 (nur KI), intraoperativ bei 0,66 (Anästhesist), 0,78 (KI und Anästhesist) bzw. 0,81 (nur KI). Beide Arbeiten zeigen eindrücklich das enorme Potenzial, mit dem KI Ärzte beim Überwachen der Vitalfunktionen unterstützen könnte, und dass viele ML-lesbare Daten im klinischen Alltag bereits digital ungenutzt zur Verfügung stehen. Kritisch für die Arbeit von Lundberg et al. [90] ist anzumerken, dass diese monozentrische Studie bisher rein retrospektiv durchgeführt wurde. Auch die Erfassung der Leistung der Anästhesisten fand nicht intraoperativ statt, sondern anhand von Diagrammen am Computer. Daher fehlt auch hier eine prospektive, klinische Evaluation mit harten Endpunkten.

Die Vermeidung von Awareness (7–11 pro 1.000 Narkosen) ist ein Hauptqualitätsmerkmal für den Patientenkomfort, jedoch mit gängigen Messungen wie dem Bispektralindex (BIS) nicht sicher möglich. Tacke et al. kombinierten Elektroenzephalogramm (EEG) und akustisch evozierte Potentiale (AEP), um Reaktionen auf Stimuli von bewusstlosen bzw. wachen Patienten zu erfassen. Hierzu wurde bei n = 40 Patienten eine Narkose durchgeführt und diese in Aufwachphasen bzw. Narkosevertiefung unterteilt. In den Wachheitsphasen wurde die Reaktion auf erteilte akusti-

sche Kommandos erfasst. In ihrer Grundlagenarbeit konnten sie so mit einer Sensitivität von bis zu 0,935 eine Wachheit vorhersagen. Mittels EEG oder AEP alleine lag die Sensitivität mit 0,91 bzw. 0,88 auf ähnlichem Niveau, jedoch deutlich höher verglichen zum BIS (Sensitivität 0,53, Spezifität 0,69) [25,26]. Die Studie zeigt einen vielversprechenden Ansatz im Sinne einer Machbarkeitsstudie, jedoch ist auch hier die Studienpopulation zu klein und es handelte sich um komplett kontrollierte Studienbedingungen. Zudem fehlen in der Arbeit Angaben zur AUC-ROC, Spezifität, PPV und NPV, sodass die Performance des Algorithmus nicht abschließend beurteilt werden kann. Ein zukünftiges Modell zur Awareness-Detektion sollte zusätzlich auch die Anästhetikagaben zu einem Gesamtmodell zusammenführen.

Eine adäquate perioperative Schmerztherapie erhöht den Patientenkomfort, hat positive Auswirkungen auf das Behandlungsergebnis und reduziert Medikamentennebenwirkungen. Ben-Israel et al. entwickelten auf Basis von Plethysmographie, Hautleitfähigkeit, Herzfrequenz und -variabilität und deren zeitlichen Verlauf mit Hilfe einer Regressionsanwendung das sog. Nozizeptionslevel (NoL). Dieses korrelierte sehr gut mit den Schmerzstimuli (R = 0,88) unter Allgemeinanästhesie und zeigte eine exzellente AUC-ROC von 0,97 [27]. Der Algorithmus wurde mehrfach extern validiert und ebenso klinisch erprobt [28,29]. Meijer et al. zeigten bei 50 Patienten aus zwei Zentren einen niedrigeren intra- und postoperativen Opioidbedarf, wenn die Schmerztherapie mittels NoL gesteuert wurde im Vergleich zur Steuerung anhand klassischer Parameter wie Herzfrequenz- und Blutdruckanstieg, was zu einer Art Paradigmenwechsel in der Narkoseführung führen könnte [30]. Der Algorithmus ist nach der FDA-Zulassung kommerziell erhältlich [31]. Für eine Übersicht über ML/KI in der Anästhesie sei auf die Arbeit von Hashimoto verwiesen [32].

Fall 2 – Notfallmedizin

Ein 40-jähriger Autofahrer verliert bei Regen und Tempo 70 km/h in einer Kurve die Kontrolle über seinen Kleinwagen und schlägt im Bereich der frontalen Beifahrerseite in einen Baum ein. Die Rettungskräfte treffen einen zentralisierten Patienten mit GCS 9, systolischem Blutdruck 90 mmHg und einer Sättigung von 90 % unter Raumluft an. Der gesamte Thorax ist druckschmerzhaft und der rechte Oberschenkel offen frakturiert. Der Notarzt entschließt sich zur präklinischen Intubation. Während des Transports kommt es zu einer Reanimation bei pulsloser elektrischer Aktivität. Durch sofortige Reanimationsmaßnahmen, Perikardtamponaden- und Pneumo-thoraxausschluss kann unter Volumengabe ein wiedereinsetzender Spontankreislauf erzielt werden. Bietet ML schon heute die Chance, Notärzte frühzeitig bei der Einschätzung von möglichen Verletzungsmustern, Therapienotwendigkeiten und bei der Reanimation zu unterstützen?

Die Vorhersagbarkeit des Zusammenhangs der Charakteristika des Autounfalls auf das Verletzungsmuster wäre für Notärzte gerade bei mehreren Fahrzeugen und Patienten wichtig für ihre Versorgungsstrategie. Kononen et al. modellierten hierzu die Auswirkungen von Geschwindigkeit, Gurtverwendung, Patientenalter, Aufprallrichtung und Fahrzeugtyp auf den **Injury Severity Score** (ISS) mit einer AUC-ROC von 0,84 [33]. Kong et al. erzielten in ihrer Kohorte mit ähnlichen Attributen eine Untertriage von 6,1 % bei einer AUC-ROC von 0,896 (Sensitivität 0,83, Spezifität 0,89) [34]. Allgemein wird eine Untertriage bei Traumapatienten von <5 % angestrebt, um zum einen die Patientensicherheit zu gewährleisten, zum anderen die Kapazität der Schockräume nicht zu überlasten. Realistische Untertriageraten betragen z. T. mehr als 10 % [35]. Daher könnten die beiden Arbeiten helfen, die Patientenversorgung zielgerichteter durchzuführen und klinische Ressourcen zu optimieren.

Die aktuelle Leitlinie zur Polytraumaversorgung empfiehlt eine invasive Atemwegssicherung bei Schock, schwerer respiratorischer Insuffizienz oder einem GCS <9 [36]. In der prähospitalen Schwerverletztenversorgung muss allerdings jederzeit mit einer Zustandsverschlechterung gerechnet werden. Die eigene Arbeitsgruppe führte mit den Daten des minimalen Notfalldatensatzes (MIND) aus dem Rettungsdienst Baden-Württemberg eine Studie mit dem Ziel der Vorhersage der Notwendigkeit der präklinischen Atemwegssicherung beim erwachsenen Polytraumapatienten durch [11]. Der MIND weist die Besonderheit auf, dass nur Vitaldaten und Scores beim Erstkontakt und bei Krankenhausübergabe erfasst werden. Nach Ausgleich eines Minderheitenproblems (Ratio 1:24) konnte eine AUC-ROC von 0,96 erzielt werden, vor allem bedingt durch die hohe Spezifität (>0,99). Sensitivität bzw. PPV lagen bei 0,73 bzw. 0,85 (AUC-PRC 0,83) und damit vergleichbar zu Modellen zur Intubationsnotwendigkeit bei Aufnahme auf die Intensivstation (z. B. Siu et al. Sensitivität 0,88; Spezifität, 0,66; AUC-ROC 0,86; PPV 0,73 [37]). Zukünftige Modelle könnten von einer Erweiterung um Echtzeit- bzw. Verlaufsparameter profitieren. Zudem müssen Strategien zur Balancierung gewählt werden, um für Patientenminderheiten eine adäquate Testgüte zu gewährleisten [11].

ML wurde für eine Vielzahl von Fragestellungen im Gebiet der prähospitalen Reanimation bspw. zur Prognoseabschätzung angewendet. Gräsner et al. errechneten die Wahrscheinlichkeit für einen wiedereinsetzenden Spontankreislauf bei prähospitaler Reanimation aus den Daten des Deutschen Reanimationsregisters (RACA-Score: **return of spontaneous circulation** (ROSC) **after cardiac arrest** [38]). Zur Anwendung kommt der Score (bestehend aus 13 standardisierten Attributen) zur Abschätzung der beobachteten und berechneten ROSC-Rate im Sinne einer Qualitätskontrolle im Deutschen Reanimationsregister. Die Diskriminierungsfähigkeit (AUC-ROC 0,71) wurde mehrfach extern validiert, konnte aber bereits von anderen Algorithmen übertroffen werden (z. B. Liu et al. AUC-ROC 0,806) [39–43].

Der Frage, ob ein neurologisch gutes Ergebnis auch bei präklinisch noch nicht erzieltem ROSC vorhergesagt werden kann und Reanimationsmaßnahmen somit nicht abgebrochen werden sollten, haben sich Seo et al. in einer koreanischen Registerstudie angenommen [44]. Dabei erzielten sie mittels Daten aus Präklinik und Schockraumversorgung (Gesamtdauer der Reanimation, Epinephringabe, Rhythmus, Defibrillation, endotracheale Intubation, mechanische Kompressionshilfe) eine AUC-ROC von 0,926. Bei hohen Sensitivitäten und Spezifitäten von über 0,85 erzielte das Modell allerdings eine PPV von nur 0,109 bzw. NPV von 0,997. Cheng et al. konnten mit ihrem registerbasierten Extremgradientenmodell zudem zeigen, dass am Ende der intensivmedizinischen Versorgung sowohl ein gutes neurologisches Ergebnis (AUC-ROC 0,956, Sensitivität/Spezifität 0,875 bzw. 0,904, PPV 0,437), die Entlassung (AUC-ROC 0,866, Sensitivität/Spezifität 0,84 bzw. 0,862, PPV 0,6) als auch die 30-Tage-Mortalität (AUC-ROC 0,831, Sensitivität/Spezifität 0,745 bzw. 0,825, PPV 0,564) bereits ausreichend gut vorhergesagt werden können [45]. Beide Arbeiten könnten nach weiterer Validierung die Basis für eine frühzeitige Einschätzung des Outcomes nach Reanimation bilden.

Um die Qualität der Reanimationsmaßnahmen zu steigern, sollten Phasen ohne Thoraxkompression minimiert werden. ML könnte helfen, defibrillierbare Rhythmen noch während der Herzdruckmassage sicher zu detektieren. Nach Filterprozessierung des EKG erzielte das künstliche neuronale Netzwerk von Isasi et al. eine Sensitivität und Spezifität von über 0,95 [46]. In einer Folgestudie zeigten sie ebenfalls, dass solch ein Algorithmus auch bei automatischen Reanimationshilfen einen defibrillierbaren Rhythmus besser erkennen konnte als die geräteeigene Software (Sensitivität 0,92, Spezifität 0,96) [47]. Beide Arbeiten haben das Potenzial zu einer substanziellen Verbesserung im klinischen Alltag und gerade auch in der Laienreanimation unter Verwendung automatischer externer Defibrillatoren. Eine prospektive Evaluation steht aber ebenso wie eine nachgewiesene Wirksamkeit noch aus. Eine Übersicht über weitere Studien zu ML und Reanimation bietet die Arbeit von Okada et al. [48].

Zukünftig könnten aus der Klinik übertragene Ansätze wie eine Hypotonievorhersage die präklinische Versorgung und ihre limitierten diagnostischen Möglichkeiten verbessern. Für Echtzeitmodellierungen bleibt jedoch die zeitnahe Datenerfassung eine Herausforderung, da die Daten normalerweise größtenteils im Einsatzverlauf nachdokumentiert werden.

Fall 3 – Intensivmedizin

Eine 71-jährige Patientin klagt auch am 4. postoperativen Tag nach elektiver Sigmaresektion bei reduziertem Allgemeinzustand weiterhin über Bauchschmerzen. In der Abdomenbildgebung ergeben sich Hinweise auf eine Anastomoseninsuffizienz. Bei zunehmender Kreislaufinsuffizienz wird die Indikation zur notfallmäßigen Laparotomie gestellt. Intraoperativ bestätigt sich der Verdacht bei zusätzlicher Vierquadrantenperitonitis. Die Patientin wird postoperativ im septischen Schock auf die Intensivstation verlegt. Hier entwickelt sie ein akutes Nieren- und Lungenversagen. Kann ML helfen, frühzeitig Krankheitszustände und Komplikationen wie Sepsis oder Organversagen zu detektieren und valide Prognosen zum Behandlungserfolg zu erstellen?

Die intensivmedizinische Sepsistherapie ist durch den hohen Grad an Digitalisierung, weltweit einheitliche Diagnosekriterien, aber auch durch bereits etablierte konventionelle Scoring-Systeme ein ideales Forschungsfeld für ML-Anwendungen. Prognostisch ist neben Kreislaufstabilisierung, Fokussanierung, antimikrobieller Therapie und potenziellem Organversagen die frühzeitige Diagnosestellung bedeutsam. Prinzipiell konnte bereits gezeigt werden, dass ML-Anwendungen mittels klinischer und laborchemischer Daten klassischen Prognose-Scores oftmals überlegen sind. Wang et al. erstellten mit Hilfe von Blutbild, Blutgasen, Elektrolyten, Leber- und Nierenwerten ein Modell zur Sepsis-Vorhersage mit einer AUC-ROC von 0,91 (Sensitivität 0,87 bzw. Spezifität 0,89) [49]. Interessanterweise kam das auf einer Kohorte chinesischer Intensivpatienten basierende Modell ohne C-reaktives Protein oder Procalcitonin aus und umfasste einige Parameter des **Sequential Organ Failure Assessment Score** (SOFA)-Score. Somit könnte eine Erweiterung des SOFA-Scores eventuell früher zur Diagnosefindung bei Sepsis beitragen, zumal dieser in einer anderen vorselektionierten Kohorte eine AUC-ROC von 0,89 bei einer Sensitivität von 0,99 bzw. Spezifität von 0,79 aufwies (PPV 0,57) [50]. Allerdings fehlt auch dieser retrospektiven Studie die externe Validation bzw. fand bereits eine Patientenvorselektion statt, sodass ein allgemeingültiges Modell noch nicht verfügbar ist.

Seymour et al. erstellten ein robustes Modell aus retrospektiven Daten, welches vier Phänotypen (α, β, γ, und δ) bei Sepsispatienten differenzierte, die sich in Inflammationsmustern und klinischem Outcome unterschieden. Die Autoren hoffen, dass weitere Studien helfen, Therapieregime und -erfolge bei diesen Untergruppen besser zu verstehen [51]. Dies wäre ein Schritt auf dem Weg zu einer individualisierten Sepsistherapie, sofern sich aus der Gruppierung auch therapeutische Implikationen ergeben würden. Einen ersten Ansatz für eine individualisierte Therapie stellt Hydrocortison im septischen Schock dar. Pirracchio et al. erstellten aus gepoolten Studien zur Hydrocortisongabe ein Modell, welches auf einer erhöhten Überlebensrate unter Hydorcortisongabe bei Sepsispatienten basierte. Der berechnete Nettonutzen für eine individualisierte Therapie lag etwa um den Faktor 3 höher, als wenn alle Patienten es erhalten hätten oder basierend auf dem SAPS II (**Simplified Acute Physiology Score**) entschieden worden wäre [52]. Somit könnte ausgerechnet ML der Diskussion um die Evidenzstärke von Hydrocortison in der Sepsistherapie neuen Schwung verleihen, sofern eine prospektive Evaluation die Ergebnisse bestätigt.

Auch zum septischen Organversagen existieren vielversprechende Arbeiten. Beim sepsisassoziierten Nierenversagen

konnten Yue et al. ein gradientengeboostetes Modell als den Algorithmus mit der höchsten Diskriminierungsrate ermitteln (AUC-ROC 0,817) [53]. Ob eine zuvor eingeleitete Volumentherapie das Nierenversagen verbessern kann, konnten Zhang et al. mit Hilfe der Attribute Serumkreatinin, Harnstoff, Alter und Serumalbumin ebenfalls zuverlässig vorhersagen (AUC-ROC 0,86) [54]. Auch die ML-basierte Vorhersage einer sepsisinduzierten Koagulopathie gelang Zhao et al. besser als mit klassischen Scores (z. B. 0,842 vs. 0,752 SIC-Score) [55]. Generell kann die 30-Tage-Mortalität für Sepsispatienten bereits zuverlässig ermittelt werden (AUC-ROC 0,80–0,876) [52]. Diese Studien machen deutlich, dass gerade in diesem Themengebiet Komplikationen und Prognose durch ML besser abschätzbar und erstere ggf. früher therapierbar sein könnten [56–58].

Beim akuten Lungenversagen (ARDS) können eine frühzeitige Diagnostik und Therapie helfen, den pathophysiologischen Verlauf zu beeinflussen. Wu et al. entwickelten ein Modell aus nicht invasiven Parametern von Überwachungsmonitoren und Beatmungsgeräten, welches ein schweres ARDS mit einer AUC-ROC von 0,869 bei einer Sensitivität von 0,61 und einer Spezifität von 0,92 vorhersagte. Statt mit dem **Lung Injury Score** (LIS) (AUC-ROC 0,82, Sensitivität 0,84, Spezifität 0,67) verglichen sie ihr Modell mit dem **Oxygenation Saturation Index** (OSI) (AUC-ROC 0,65), der jedoch nicht für eine Vorhersage des ARDS erstellt wurde, sondern für Mortalität und Ventilationsdauer [59, 60]. Interessanterweise war die Diskriminierungsfähigkeit der Modelle in der Frühphase bis zwei Stunden nach Krankenhausaufnahme am höchsten, weswegen die Autoren ihr Modell auch möglicherweise für die Präklinik als geeignet ansahen [61]. Da jedoch die ermittelte Sensitivität gering erscheint, läge ein zukünftiger Ansatz vielleicht in einer Kombination des Modells mit dem LIS. Um zwischen einem hyper- und hypoinflammatorischen ARDS zu differenzieren, nutzten Sinha et al. klinische Daten und Werte aus Routinelaboranalysen aus vier großen ARDS-Studienkollektiven. Daraus resultierte ein Modell mit einer AUC-ROC 0,94 [62]. Bai et al. fanden in ihrer Untersuchung zum sepsisassoziierten ARDS in ihrer Kohorte drei Cluster (hypo-, hyperinflammatorisch und chronisch), welche sich in der Mortalität in Abhängigkeit von den gewählten Therapieformen wie den Beatmungsparametern unterschieden. Generell konnte ein ARDS mit einer AUC-ROC von 0,895 differenziert werden [63]. Die Arbeiten verdeutlichen, dass viele Prozesse beim ARDS noch unverstanden, prinzipiell allerdings individuell abschätzbar sind. Daraus weitergehende therapeutische Implikationen abzuleiten, wäre im Moment jedoch noch verfrüht [64].

Fall 4 – Schmerz- und Palliativmedizin

Ein 76-jähriger Mann stellt sich bei einem Schmerztherapeuten wegen einer Post-Zoster-Neuralgie im Bereich Th4–6 rechts vor. Die hausärztliche Schmerztherapie mittels Opioidtherapie sei unzureichend. An Vorerkrankungen bringt der Patient einen arteriellen Hypertonus, Nikotinabusus sowie ein inoperables Bronchialkarzinom mit, welches sich im Staging nach kürzlicher Chemotherapie weiterhin progredient zeigte. Kann maschinelles Lernen die Therapie einer Post-Zoster-Neuralgie verbessern? Welches Potenzial liegt in der palliativmedizinischen Versorgung und bei Entscheidungen am Lebensende?

Therapierefraktäre, chronische neuropathische Schmerzen gehören zu den schwierigsten zu therapierenden Schmerzsyndromen und schränken die Lebensqualität stark ein. Deren Auftreten kann durch ML im speziellen Fall der Post-Zoster-Neuralgie vorhergesagt werden [65]. Dabei wurden eine subakute Infektion, schwere Läsionen, Depression und Hypertension als Risikofaktoren identifiziert und eine sehr gute Klassendiskriminierung mit einer AUC-ROC 0,918 erzielt. Die Ergebnisse sollten für eine klinisch prospektive Studie genutzt werden, um frühzeitig das Entstehen eines Schmerzgedächtnisses zu verhindern.

Mit Blick auf eine individualisierte Therapie gerade unter dem Eindruck der Opioidkrise gelang es Gudin et al., die Wirksamkeit von topischen Analgetika bei chronischen Schmerzpatienten unter Opioiddauertherapie zum Zwecke der Dosisminderung und Lebensqualitätssteigerung vorherzusagen. Ihre Modelle erzielten eine AUC-ROC von 0,73–0,87, indem sie Verlaufsveränderungen im **Brief Pain Inventory**-Schmerzfragebogen erfassten. Nach Implementierung in ein klinikinternes Tool wurde ein Therapieversuch mit topischen Analgetika danach gesteuert (94 % Erfolgsrate) [66]. Diese Studie ist insbesondere aufgrund der prospektiven, internen Validierung des Tools und weiteren Verwendung hervorzuheben.

Um die Versorgung am Lebensende zu verbessern, entwickelten Soltani et al. eine KI, die sowohl individuell als auch populationsbasiert medizinische, pflegerische, soziale und psychologische Belange vorhersagen konnte [67]. Hierzu verbanden sie Entlassmanagementdaten wie Wohnort und Diagnosen eines nationalen iranischen Krebszentrums, um ein selbstlernendes neuronales Netzwerk zu bilden. Es gelang ihnen so, eine verbesserte individuelle und populationsbasierte Versorgung durch optimierte zeitliche und örtliche Ressourcensteuerung zu erzielen. Sandham et al. nutzten maschinelles Lernen, um aus verschiedenen palliativmedizinischen Skalen Attribute herauszufiltern, die eine stabile von einer instabilen, sich verschlechternden oder terminalen Phase unterscheiden [68]. Der Erfolg der Modelle war in den einzelnen Phasen leider inkonstant, die identifizierten Attribute wie Müdigkeit, Kurzatmigkeit oder Übelkeit/Appetitverlust aber prinzipiell digital für eine zukünftig verbesserte Versorgung erfassbar. Wang et al. ermittelten indirekt über eine zeitnahe Mortalitätsvorhersage bei dementen Patienten deren Bedarf für eine palliative Therapie am Lebensende. Mittels automatisierter Texterkennung aus Akteneinträgen konnten sie u. a. die Attribute Delir, Krebs, Schmerz, Arthritis, Ernährungsstatus, Schluckstörung oder häufige Arztkonsultation identifizieren, die die

Mortalität als Surrogatparameter für eine mögliche palliativmedizinische Versorgung vorhersagten (AUC-ROC >0,94) [69]. Guo et al. modellierten für Patienten mit palliativer Chemotherapie bei Bronchialkarzinom eine AUC-ROC von 0,897 für die Vorhersage einer pulmonalen Infektion [70]. Als Risikofaktoren identifizierten sie Alter ≥60 Jahre, einen Krankenhausaufenthalt ≥14 Tage, kombinierte Chemotherapie, Myelosuppression, chirurgische Eingriffe, Diabetes und Hormonsubstitution. Diese vielversprechenden verschiedenen Ansätze zur Verbesserung der palliativmedizinischen Versorgung bedürfen jedoch einer Adaptation an lokale Gegebenheiten sowie einer Reevaluation.

Beispielhaft publizierten Meier et al. in einer Machbarkeitsstudie ein Modell für ethische Entscheidungswege am Lebensende. Als Datenquelle wählten sie 69 Berichte der klinischen Ethikkommission. Ihre ML-basierte kognitive Karte erfasste nicht nur die ethische Grundhaltung, sondern berücksichtigte auch Einwilligungsfähigkeit, Alter, Patientenzustand, niedergelegter bzw. mutmaßlicher Patientenwille sowie aus medizinischer Sicht die Handlungsoptionen des Nichtschadens bzw. der Wohltat. Letzteren Handlungsoptionen wohnt ein individuell-situativer Konflikt zwischen Zugewinn an Lebenstagen oder Lebensqualität inne. Unter Berücksichtigung aller Punkte errechnet der Algorithmus über die Zeit die Wahrscheinlichkeit, durch welche Art die endgültige Entscheidung maßgelblich beeinflusst wird und ob dies dem Prinzip des Nichtschadens bzw. der Wohltat entspricht. Er erzielte Übereinstimmungen zu der von den Autoren ermittelten ethischen Position und der tatsächlichen Position der Ethikkommission von 75–92 %. Als Limitation war die geringe Fallzahl zu nennen, sowie dass bestimmte Prinzipien wie Gerechtigkeit (z. B. die Allokation bestimmter Maßnahmen in Mangelsituationen) nicht berücksichtigt werden konnten [16]. Die Studie zeigt exemplarisch die Herausforderungen ethischer Fragestellungen an KI-Anwendungen. Zudem werden mögliche Konfliktfelder wie Fremdbestimmung am Lebensende, unzureichende Berücksichtigung individueller ethischer Standpunkte, Angst vor Diskriminierung oder der Umgang mit divergierenden Entscheidungen sichtbar [71]. Die Erstellung von Algorithmen zur Unterstützung ethischer Entscheidungen stellt eine große Herausforderung dar, zumal die Frage, welchen Einfluss KI auf Entscheidungen am Lebensende haben sollte, einer breiten gesellschaftlichen Diskussion bedarf [72].

Diskussion

Zwar zeigen die exemplarischen ML-Studien enormes Potenzial in der Forschung und alltäglichen Patientenversorgung, jedoch stehen Entwicklung und Praxistauglichkeit vieler Anwendungen noch am Anfang. Die FDA hat bis jetzt für die AINSP lediglich einige wenige Anwendungen zugelassen, hierunter die HPI-Software, den NoL-Algorithmus, ein Vorhersagemodell zur zentralvenösen Sättigung sowie EKG-Analyse-Algorithmen (Stand 11/2023 [31]). Doch warum liegen Potenzial, publizierte und zugelassene Modelle so weit auseinander?

Entwicklungsprozess

In einer Übersichtsarbeit von van de Sande et al. in der Intensivmedizin konnten die Autoren aus 494 Studien im Jahr 2021 noch keine Anwendung identifizieren, die den Weg in den klinischen Alltag gefunden hatte [73]. Um den Erwartungen an die Möglichkeiten von KI-Anwendungen gerecht zu werden, sei laut Autoren daher ein strukturiertes Vorgehen bei der Erstellung und Implementierung solcher Programme nötig [74]. Hierzu existieren mehrere Entwürfe, z. B. ein von Fleuren et al. an die Raumfahrt adaptiertes neunstufiges Konzept (Tab. 2) [75]. Van de Sande et al. schlugen analog zur Arzneimittel-

Tabelle 2
Einsatzfähigkeit und Entwicklungsprozess von Anwendungen des maschinellen Lernens in der Medizin nach der modifizierten NASA-Definition, angelehnt an [75].

Level	NASA-Definition	Klinische Definition	Erklärung
1	Problemobservation und -beschreibung	Klinische Problemidentifikation	Literaturrecherche, die Verbesserungspotential bei KI-Anwendung ergibt
2	Beschreibung eines Technologiekonzepts	Lösungsvorschlag	Projektskizzierung und -registrierung
3	Analytische und experimentelle kritische Funktion und/oder Proof of Concept	Modellentwicklung I	Machbarkeitsstudien durch prototypische Modelle, die ein Potential für Vorhersagewahrscheinlichkeit bieten, Modelloptimierung und Validierung an internen Datenquellen
4	Komponententestung im Labor	Modellentwicklung II	
5	Komponententestung im erweiterten Modell	Modellvalidierung	Externe Validierung an retrospektiven oder prospektiven Daten
6	Komponententestung im erweiterten Modell	Echtzeittestung	Prospektive Testung mit Echtzeitdaten im Vergleich zur Standardbehandlung
7	Systemprototyp in der Arbeitsumgebung	Integration in den Arbeitsprozess	Phase-2-Studie zur Anwender- und Patientensicherheit
8	Durch Test und Demonstration aller Komponenten geprüftes, vollständiges System	Outcome-Evaluation	Phase-3-Studie, doppelblind randomisiertes Design
9	Endgültig getestetes System durch Missionsoperation	Integration in den Arbeitsalltag in multiplen Zentren	Langzeitbeobachtung zu Sicherheit, Verhaltensänderungen, Anpassung an verändertes Verhalten/Therapiekonzepte

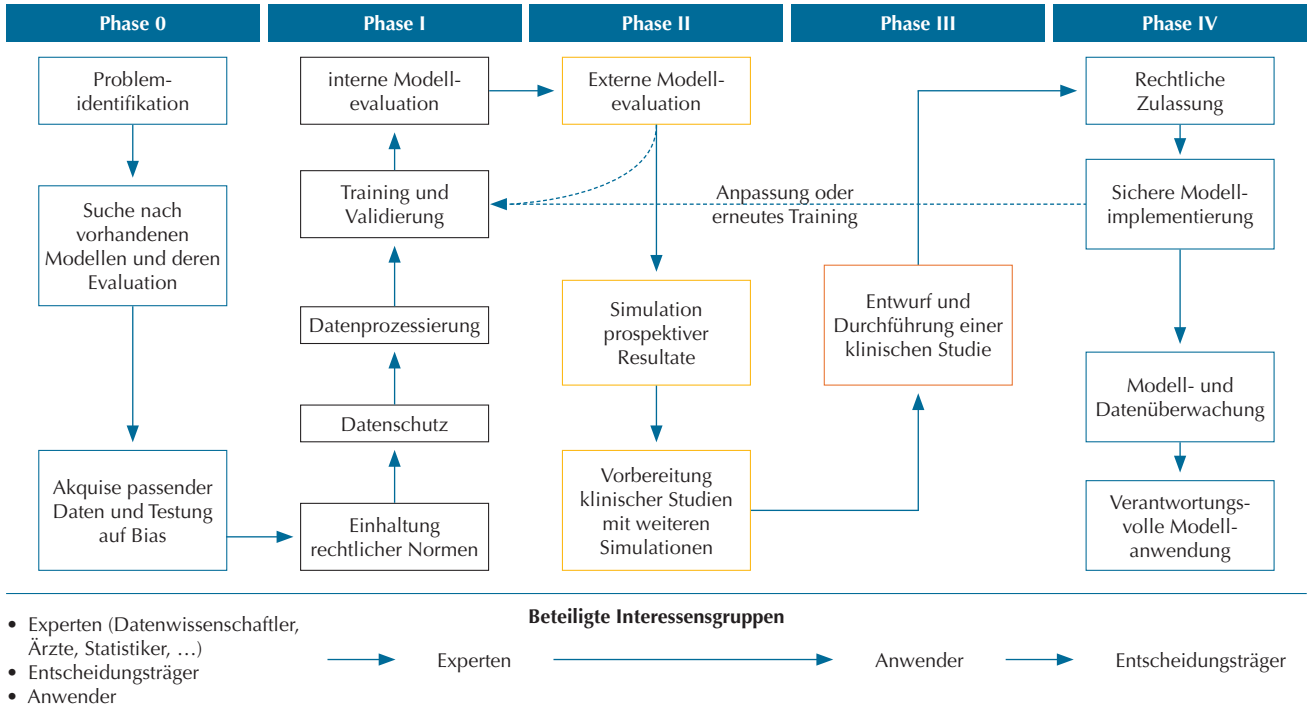
NASA: Amerikanische National Aeronautics and Space Administration.

zulassung vier Phasen von der (Daten-) Vorbereitung über Modellierung und die interne und externe Testung bis hin zur Implementierung mit Nachüberwachung vor (Abb. 1) [74]. Unter Berücksichtigung beider Ansätze sind die Gründe, warum im überwiegenden Fall noch keine Praxisfähigkeit vorliegt, ersichtlich und vielschichtig. In einem ersten Schritt muss ein relevantes Themenfeld identifiziert werden, welches eine Datengrundlage für ML aufweist und für das bisher keine oder nur unzureichende Modelle existieren. Im Rahmen der Modellentwicklung sind im zweiten Schritt vor allem die Datengrundlage und -analyse zu nennen. Zum Zeitpunkt der internen Testung bietet sich zudem auch der Vergleich mit einem existierenden Scoring-System an. In einem dritten Schritt sollte das Modell extern validiert werden mit dem Ziel, genug Sicherheit für eine klinisch-prospektive Studie zu generieren. Diese stellt den vierten Schritt dar. Exemplarisch hierzu sei die Evaluation des **Epic Sepsis**-Modells (ESM) durch Wong et al. genannt. Das ESM wurde in den 2010er-Jahren an drei US-amerikanischen Standorten an über 400.000 Patienten erstellt und landesweit ohne genaue Nennung der Testgütekriterien vermarktet und eingesetzt. Die Autoren validierten nun an ihrer Klinik das Programm retrospektiv verglichen mit der klinischen Alltagspraxis. Dabei zeigte sich für das Auftreten einer Sepsis eine AUC-ROC von lediglich 0,63 sowie ein PPV von lediglich 0,12 bei einer Sensitivität von 0,33 (Spezifität 0,83). Die Autoren mahnten eine bessere Validierung und Kalibrierung vor einer nationalen Verbreitung von KI-Programmen an und wiesen zudem auf eine mögliche Anwenderbelastung durch falsche Alarmierungen hin [76]. In der fünften Phase muss das Modell in den breiten klinischen Alltag integriert werden, nachdem es eine behördliche Zulassung [77,78] erhalten hat und nachdem sichergestellt wurde, dass eine kontinuierliche Überwachung der weiteren Testgüte z. B. bei sich verändernden Leitlinien sichergestellt ist. Dieser langwierige Erstellungs- und Zulassungsprozess wirft zudem die Frage nach der Modellentwicklung in nicht wirtschaftlich gewinnversprechenden Themengebieten auf.

Interaktion KI – Mensch

Gerade im alltäglichen Umgang mit KI durch Anwender sind einige Probleme bisher noch nicht ausreichend untersucht. Ziel muss der kritische Anwender sein, der KI-Anwendungen als einen Baustein des Diagnose- und Therapieprozesses versteht. Ein sorgloses Vertrauen kann sonst zu einer potenziellen Patientengefährdung führen, ebenso wie die vorschnelle Ablehnung dieser Methoden [79]. Die Mensch-KI-Interaktion ist ein sich gerade entwickelndes Forschungsfeld. Es wird durch Themen wie Vertrauen (in die KI ohne Kennzeichen der vertrauensvollen Mensch-Mensch-Interaktion), Reduktion/Vergrößerung

Abbildung 1



Übersicht über die Entwicklung und Implementierung von KI-Anwendungen sowie die beteiligten Interessensgruppen nach van de Sande et al. [74]. Nur ein erfolgreicher Abschluss einer Phase sollte zur Eröffnung der nächsten führen. Gegebenenfalls muss das Modell wieder angepasst oder neu trainiert werden.

von Unsicherheiten (KI gibt nach Wahrscheinlichkeitsabgleich eine definitive Antwort) und Umgang mit kognitiven Bias (z. B. Autoritäts-Bias, Status-quo-Bias) charakterisiert [79]. Hierzu existieren einige Lösungsvorschläge, die sowohl von Seiten der Hersteller als auch durch die Anwender selbst aufgegriffen werden können. Die Hersteller könnten den zugelassenen Anwendungen eine Art „digitalen Beipackzettel" beifügen, der ihre wichtigsten Eigenschaften wie Anwendungsgebiet, Testkohorte, Balancierung, Algorithmus, Gütekriterien, Validierung und Limitationen kurz, prägnant und jederzeit schnell ersichtlich auflistet. Als Präsentationsform für die Ergebnisse kann beispielhaft eine wertebewusste Darstellung (**value sensitive design** [80]) verwendet werden, die Ergebnisse, beabsichtigtes Ziel, den Gewichtungseffekt bestimmter Ergebnisse hierauf sowie die Anwenderansprüche und -wertvorstellungen in der optischen Darstellung berücksichtigt. Die Anwender wiederum sollten bereits in der klinischen Ausbildung eine Wissensvermittlung über die Grundzüge von Algorithmen sowie von Testgütekriterien erhalten, da die Verantwortung der Testinterpretation sowie der abgeleiteten Therapieentscheidungen bei ihnen liegt. Im Rahmen einer Implementierung in klinikeigene Systeme sollte zudem eine Nutzerschulung erfolgen. Auch sind hier Nachweise analog zu Einweisungen von Medizinprodukten denkbar. Die Nutzer müssen zudem informiert und dahingehend motiviert werden, einen ggf. gerechtfertigten Mehraufwand an Dokumentation zu akzeptieren, um an anderer Stelle Arbeitserleichterungen oder eine verbesserte Behandlungsqualität zu erzielen.

Implikationen für Medizin und Politik

Wie zuvor gezeigt, sind viele Studien oftmals noch Grundlagenarbeiten, deren Auswirkungen oder Erkenntnisse weitenteils noch nicht absehbar sind. Daher sollte der Schwerpunkt weiterer Forschungsaktivität nicht allein darin bestehen, immer neue Modelle mit möglichst guter Trennschärfe zu generieren. Nach Meinung der Autoren sind Validierungs- und Replikationsstudien für den weiteren Fortschritt bei der Translation von ML-Anwendungen in die Klinik ebenso wichtig. Außerdem besteht großer Bedarf an Arbeiten, die zu einer Verbesserung der Verfügbarkeit und Qualität der Datengrundlagen, der Methoden für die Modellierung und praktische Testung sowie der Implementierung von ML-basierten Anwendungen führen. Das Einhalten von Qualitätsstandards ist selbstverständlich auch bei diesen zu fordern. Hierbei ist zum einen die standardisierte Auswertung und Berichterstattung analog des TRIPOD-Schemas (**Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis**) zu nennen, zum anderen das PROBAST-Hilfsmittel (**Prediction model Risk Of Bias ASsessment Tool**), welches Verzerrungen bei der Entstehung und Validierung von ML-Modellen aufzeigen soll [81,82]. Eine Erweiterung beider Standards hin zu KI-Modellen ist in der Entstehung [83]. Die flächendeckende Veröffentlichung von Quellcode und Datensätzen und der Aufbau multiethnischer Datenbanken könnten den Entwicklungs- und Validierungsprozess zudem beschleunigen. Es liegt auch an Ärzteschaft und Politik, zum einen eine Dateninfrastruktur von der Präklinik bis hin zur Entlassung/Rehabilitation aufzubauen, um das Potenzial von KI-Anwendungen ausschöpfen zu können. Außerdem müssen ethische und juristische Regeln definiert werden, um Patientensicherheit auf der einen und Therapiequalität auf der anderen Seite zu gewährleisten. Hierzu existieren bspw. eine Stellungnahme des deutschen Ethikrates oder ein Gesetzgebungsverfahren der Europäischen Union [84,85]. Ziel muss es sein, dass KI-Anwendungen bezogen auf Wohltätigkeit, Transparenz, Nichtschaden, Autonomie, Gerechtigkeit und Datenschutz ethisch vertretbare Handlungsunterstützung anbieten und diese von Patienten und Anwendern sicher nachvollzogen werden können [71]. Für Patienten darf nicht der Eindruck entstehen, einer intransparenten Software ohne menschliche Kontrollinstanz hilflos ausgeliefert zu sein.

Trotz des aufwendigen Entwicklungsprozesses sollten sich auch rein klinisch tätige Anästhesisten schon frühzeitig mit dem Themenkomplex befassen, um nicht nur Güte und Entwicklungsstand eines Modells bzw. einer Anwendung analog zu Entwicklung von Arzneimitteln und Medizinprodukten einschätzen zu können. Zudem kann die Auseinandersetzung mit dem medizinischen Inhalt, insbesondere den in den Modellen verwendeten Attributen, den klinischen Blick schärfen. Ebenso bewirkt eine Auseinandersetzung mit bisher weniger beachteten Themengebieten oftmals bereits eine Sensibilisierung hierfür [4]. Außerdem ermöglichen KI-Anwendungen eine neuartige Konsultationsform des Fachwissens anderer Disziplinen. Du-Harpur et al. beschreiben exemplarisch für KI-Anwendungen in der Dermatologie zum Melanom-Screening, wie solch ein Hilfsmittel von Ärzten anderer Fachdisziplinen, aber auch von Laien verwendet werden kann. Während Ärzte ihre bisherigen Behandlungskonzepte hiermit ergänzen und diese ggf. nochmals mit Experten rücksprechen können, kann der breite Einsatz durch Laien zu einer Übertherapie führen [86]. Auch andere Gesundheitsfachberufe wie Notfallsanitäter oder Pflegeberufe werden KI-Anwendungen verwenden. Welche Auswirkungen dies auf Zusammenarbeit, Wissensgenerierung und Kommunikation innerhalb eines Teams respektive in der Arzt-Patienten-Beziehung haben wird, ist Gegenstand aktueller Forschung [87].

Limitationen

Limitationen der Studien liegen in der exemplarischen Literaturauswahl, welche für die Verdeutlichung von Chancen, Risiken und Problemen bei der Erstellung und Validierung von KI-Modellen selektiert wurde. Allein eine vollständige Übersicht eines Kerngebietes würde über den Umfang der vorliegenden Untersuchung hinausgehen. Ferner geht der Artikel nicht weiter auf mögliche soziale und wirtschaftliche Aspekte von KI ein wie Umstrukturierung der Arbeitsprozesse, Kosten-Nutzen-Analysen respektive eine Schwächung der mensch-

lichen Position im Alltag durch eine KI-Dominanz, kombiniert mit einer Entfremdung von der Gesellschaft (**„human enfeeblement“**) [5]. Zudem wurde auch der potenziell missbräuchliche Einsatz von KI nicht weiter diskutiert, sei es durch manipulative KIs (z. B. um Patienten zu bestimmten Therapien zu drängen), Deepfakes und Hacking (z. B. in der Telemedizin), abtrünnige KI (**„Rogue AI“**, z. B. Beleidigungen durch Chatbots) oder die Anwendung von ML für militärische und bioterroristische Zwecke [1,88,89]. Um diesen Gefahren zu begegnen, ist eine begleitende, umfassende Kontrolle von KI wie oben bereits skizziert unabdingbar.

Schlussfolgerung

In der AINSP besteht großes Potenzial für ML- und KI-Anwendungen. Internationale Studien legen eine Ausweitung der Anwendungsgebiete nahe. Daher ist ein Grundverständnis hierüber unabdingbar. Bevor jedoch Anwendungen implementiert werden, sollten für extern validierte Programme nach ggf. lokaler Anpassung auch eine Anwenderschulung und Nachbeobachtung erfolgen. Anwender und Patienten müssen sicher sein, dass rechtlich zugelassene KI-Anwendungen nicht gegen ethische Prinzipien wie Wohltätigkeit, Transparenz, Nichtschaden, Autonomie, Gerechtigkeit und Datenschutz verstoßen. Maschinelles Lernen könnte somit bald vor allem als Unterstützungstool in der Diagnostik und als Behandlungskonzept sinnvoll eingesetzt werden. Die Rolle des Arztes wird in der Ergebnisinterpretation und Therapieentscheidung liegen. Die handwerklichen Fähigkeiten, Erfahrung, Empathie, Gesprächsführung, ganzheitliche Betreuung und Behandlung bleiben auf nicht absehbare Zeit die unersetzliche Kernkompetenz aller pflegerischen und ärztlichen Berufsgruppen.

Korrespondenzadresse

Dr. med.
André Luckscheiter

Klinik für Anästhesiologie, Operative Intensivmedizin und Notfallmedizin
Klinikum Ludwigshafen
Bremserstraße 79
67063 Ludwigshafen am Rhein,
Deutschland

Tel.: 0621 503 3000
Fax: 0621 503 30024

E-Mail: luckscha@klilu.de

ORCID-ID: 0000-0002-5724-7130

Literatur

1. Patwardhan A. Artificial Intelligence: First Do the Long Overdue Doable. J Prim Care Community Health 2023;14:21501319231179559
2. Kagerbauer S, Blobner M, Ulm B, Jungwirth B: Die Zukunft hat schon begonnen. Wie maschinelles Lernen Anästhesie und Intensivmedizin prägt. Anästh Intensivmed 2020;61:85–96
3. Peine A, Lütge C, Poszler F, Celi LA, Schöffski O, Marx G: Künstliche Intelligenz und maschinelles Lernen in der intensivmedizinischen Forschung und klinischen Anwendung. Anästh Intensivmed 2020:372–384
4. Witten IH, Eibe F, Hall MA, Pal CJ: Data Mining: Practical Machine Learning Tools and Techniques. 4th Edition. Cambridge (MA), United States: Elsevier; 2016
5. Doya K, Ema A, Kitano H, Sakagami M, Russell S: Social impact and governance of AI and neurotechnologies. Neural Netw 2022;152:542–554
6. Connor CW: Artificial Intelligence and Machine Learning in Anesthesiology. Anesthesiology 2019;131:1346–59
7. Kagerbauer S, Dohmen S, Reyle-Hahn S, Balzer F, Brodowski C, Ulm B et al: Weltmeister im Schneckentempo – eine Umfrage zum Status quo der Digitalisierung in Anästhesie und Intensivmedizin. Anästh Intensivmed 2023:003–013
8. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H: eDoctor: machine learning and the future of medicine. J Intern Med 2018;284:603–619
9. Sassenscheidt J, Jungwirth B, Kubitz JC: „Machine learning“ in der Anästhesiologie. Anaesthesist 2020;69:535–543
10. Nahm FS: Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol 2022;75:25–36
11. Luckscheiter A, Zink W, Lohs T, Eisenberger J, Thiel M, Viergutz T: Machine learning for the prediction of preclinical airway management in injured patients: a registry-based trial. Clin Exp Emerg Med 2022;9:304–313
12. Ozenne B, Subtil F, Maucort-Boulch D: The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015;68:855–859
13. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432
14. Chicco D, Tötsch N, Jurman G: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021;14:13
15. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA: The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2:e489–492
16. Meier LJ, Hein A, Diepold K, Buyx A: Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept. Am J Bioeth 2022;22:4–20
17. Sonntagbauer M, Haar M, Kluge S: Künstliche Intelligenz: Wie werden ChatGPT und andere KI-Anwendungen unseren ärztlichen Alltag verändern? Med Klin Intensivmed Notfmed 2023;118:366–371
18. Roth D, Pace NL, Lee A, Hovhannisyan K, Warenits AM, Arrich J, et al: Airway physical examination tests for detection of difficult airway management in apparently normal adult patients. Cochrane Database Syst Rev 2018;5:CD008874
19. Tavolara TE, Gurcan MN, Segal S, Niazi MKK: Identification of difficult to intubate patients from frontal face images using an ensemble of deep learning models. Comput Biol Med 2021;136:104737
20. Bijker JB, Persoon S, Peelen LM, Moons KGM, Kalkman CJ, Kappelle LJ, et al: Intraoperative Hypotension and Perioperative Ischemic Stroke after General Surgery: A Nested Case-control Study. Anesthesiology 2012;116:658–664
21. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al: Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. JAMA 2020;323:1052–1060
22. Li W, Hu Z, Yuan Y, Liu J, Li K: Effect of hypotension prediction index in the prevention of intraoperative hypotension during noncardiac surgery: A systematic review. J Clin Anesth 2022;83:110981
23. Schneck E, Schulte D, Habig L, Ruhrmann S, Edinger F, Markmann M, et al: Hypotension Prediction Index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: a single centre feasibility randomised blinded prospective interventional trial. J Clin Monit Comput 2020;34:1149–1158
24. Enevoldsen J, Vistisen ST: Performance of the Hypotension Prediction Index May Be Overestimated Due to Selection Bias. Anesthesiology 2022;137:283–289
25. Russell IF: The ability of bispectral index to detect intra-operative wakefulness during isoflurane/air anaesthesia, compared with the isolated forearm technique. Anaesthesia 2013;68:1010–1020
26. Tacke M, Kochs EF, Mueller M, Kramer S, Jordan D, Schneider G: Machine learning for a combined electroencephalographic anesthesia index to detect awareness under anesthesia. PLoS One 2020;15:e0238249
27. Ben-Israel N, Kliger M, Zuckerman G, Katz Y, Edry R: Monitoring the nociception level: a multi-parameter approach. J Clin Monit Comput 2013;27:659–668
28. Edry R, Recea V, Dikust Y, Sessler DI: Preliminary Intraoperative Validation of the Nociception Level Index: A Noninvasive Nociception Monitor. Anesthesiology 2016;125:193–203
29. Martini CH, Boon M, Broens SJL, Hekkelman EF, Oudhoff LA, Buddeke AW, et al: Ability of the nociception level, a multiparameter composite of autonomic signals, to detect noxious stimuli during propofol-remifentanil anesthesia. Anesthesiology 2015;123:524–534
30. Meijer F, Honing M, Roor T, Toet S, Calis P, Olofsen E, et al: Reduced postoperative pain using Nociception Level-guided fentanyl dosing during sevoflurane anaesthesia: a randomised controlled trial. Br J Anaesth 2020;125:1070–1078
31. Food and Druck Administration (FDA): Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA 2022. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (Zugriffsdatum: 15.11.2023)
32. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G: Artificial Intelligence in Anesthesiology: Current Techniquees, Clinical Applications, and Limitations. Anesthesiology 2020;132:379–394
33. Kononen DW, Flannagan CAC, Wang SC: Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accid Anal Prev 2011;43:112–122

34. Kong JS, Lee KH, Kim OH, Lee HY, Kang CY, Choi D, et al: Machine learning-based injury severity prediction of level 1 trauma center enrolled patients associated with car-to-car crashes in Korea. Comput Biol Med 2023;153:106393
35. Jeppesen E, Cuevas-Østrem M, Gram-Knutsen C, Uleberg O: Undertriage in trauma: an ignored quality indicator? Scand J Trauma Resusc Emerg Med 2020;28:34
36. Deutsche Gesellschaft für Unfallchirurgie e.V.: S3-Leitlinie Polytrauma/Schwerverletzten-Behandlung (AWMF Registernummer 187-023), Version 4.1 (31.12.2022). https://www.awmf.org/leitlinien/detail/ll/187-023.html (Zugriffsdatum: 05.03.2023)
37. Siu BMK, Kwak GH, Ling L, Hui P: Predicting the need for intubation in the first 24 h after critical care admission using machine learning approaches. Sci Rep 2020;10:20931
38. Gräsner JT, Meybohm P, Lefering R, Wnent J, Bahr J, Messelken M, et al: ROSC after cardiac arrest—the RACA score to predict outcome after out-of-hospital cardiac arrest. Eur Heart J 2011;32:1649–1656
39. Caputo ML, Baldi E, Savastano S, Burkart R, Benvenuti C, Klersy C, et al: Validation of the return of spontaneous circulation after cardiac arrest (RACA) score in two different national territories. Resuscitation 2019;134:62–68
40. Gamberini L, Tartivita CN, Guarnera M, Allegri D, Baroncini S, Scquizzato T, et al: External validation and insights about the calibration of the return of spontaneous circulation after cardiac arrest (RACA) score. Resusc Plus 2022;10:100225
41. Kupari P, Skrifvars M, Kuisma M: External validation of the ROSC after cardiac arrest (RACA) score in a physician staffed emergency medical service system. Scand J Trauma Resusc Emerg Med 2017;25:34
42. Liu N, Ong MEH, Ho AFW, Pek PP, Lu TC, Khruekarnchana P, et al: Validation of the ROSC after cardiac arrest (RACA) score in Pan-Asian out-of-hospital cardiac arrest patients. Resuscitation 2020;149:53–59
43. Liu N, Liu M, Chen X, Ning Y, Lee JW, Siddiqui FJ, et al: Development and validation of an interpretable prehospital return of spontaneous circulation (P-ROSC) score for patients with out-of-hospital cardiac arrest using machine learning: A retrospective study. EClinicalMedicine 2022;48:101422
44. Seo DW, Yi H, Bae HJ, Kim YJ, Sohn CH, Ahn S, et al: Prediction of Neurologically Intact Survival in Cardiac Arrest Patients without Pre-Hospital Return of Spontaneous Circulation: Machine Learning Approach. J Clin Med 2021;10:1089
45. Cheng CY, Chiu IM, Zeng WH, Tsai CM, Lin CR: Machine Learning Models for Survival and Neurological Outcome Prediction of Out-of-Hospital Cardiac Arrest Patients. Biomed Res Int 2021;2021:1–8; DOI: 10.1155/2021/9590131
46. Isasi I, Irusta U, Aramendi E, Eftestøl T, Kramer-Johansen J, Wik L: Rhythm Analysis during Cardiopulmonary Resuscitation Using Convolutional Neural Networks. Entropy 2020;22:595
47. Isasi I, Irusta U, Aramendi E, Olsen JA, Wik L: Shock decision algorithm for use during load distributing band cardio-pulmonary resuscitation. Resuscitation 2021;165:93–100
48. Okada Y, Mertens M, Liu N, Lam SSW, Ong MEH: AI and machine learning in resuscitation: Ongoing research, new concepts, and key challenges. Resusc Plus 2023;15:100435
49. Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, et al: A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients. Front Public Health 2021;9:754348
50. Toker A, Kose S, Turken M: Comparison of SOFA Score, SIRS, qSOFA, and qSOFA + L Criteria in the Diagnosis and Prognosis of Sepsis. Eurasian J Med 2021;53:40–47
51. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al: Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA 2019;321:2003–2017
52. Pirracchio R, Hubbard A, Sprung CL, Chevret S, Annane D; Rapid Recognition of Corticosteroid Resistant or Sensitive Sepsis (RECORDS) Collaborators: Assessment of Machine Learning to Estimate the Individual Treatment Effect of Corticosteroids in Septic Shock. JAMA Netw Open 2020;3:e2029050
53. Yue S, Li S, Huang X, Liu J, Hou X, Zhao Y, et al: Machine learning for the prediction of acute kidney injury in patients with sepsis. J Transl Med 2022;20:215
54. Zhang Z, Ho KM, Hong Y: Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019;23:112
55. Zhao QY, Liu LP, Luo JC, Luo YW, Wang H, Zhang YJ, et al: A Machine-Learning Approach for Dynamic Prediction of Sepsis-Induced Coagulopathy in Critically Ill Patients With Sepsis. Front Med 2021;7:637434
56. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al: Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46:383–400
57. Komorowski M, Green A, Tatham KC, Seymour C, Antcliffe D: Sepsis biomarkers and diagnostic tools with a focus on machine learning. EBioMedicine 2022;86:104394
58. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, et al: Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 2020;26:584–595
59. DesPrez K, McNeil JB, Wang C, Bastarache JA, Shaver CM, Ware LB: Oxygenation Saturation Index Predicts Clinical Outcomes in ARDS. Chest 2017;152:1151–1158
60. Kim BK, Kim S, Kim CY, Kim YJ, Lee SH, Cha JH, et al: Predictive Role of Lung Injury Prediction Score in the Development of Acute Respiratory Distress Syndrome in Korea. Yonsei Med J 2021;62:417–423
61. Wu W, Wang Y, Tang J, Yu M, Yuan J, Zhang G: Developing and evaluating a machine-learning-based algorithm to predict the incidence and severity of ARDS with continuous non-invasive parameters from ordinary monitors and ventilators. Comput Methods Programs Biomed 2023;230:107328
62. Sinha P, Churpek MM, Calfee CS: Machine Learning Classifier Models Can Identify Acute Respiratory Distress Syndrome Phenotypes Using Readily Available Clinical Data. Am J Respir Crit Care Med 2020;202:996–1004
63. Bai Y, Xia J, Huang X, Chen S, Zhan Q: Using machine learning for the early prediction of sepsis-associated ARDS in the ICU and identification of clinical phenotypes with differential responses to treatment. Front Physiol 2022;13:1050849
64. Bhattarai S, Gupta A, Ali E, Ali M, Riad M, Adhikari P, et al: Can Big Data and Machine Learning Improve Our Understanding of Acute Respiratory Distress Syndrome? Cureus 2021;13:e13529
65. Zhou J, Sun W, Liu Y, Yang S, Wu S, Wang S, et al: Clinical Characteristics,

Treatment Effectiveness, and Predictors of Response to Pharmacotherapeutic Interventions Among Patients with Herpetic-Related Neuralgia: A Retrospective Analysis. Pain Ther 2021;10:1511–1522
66. Gudin J, Mavroudi S, Korfiati A, Theofilatos K, Dietze D, Hurwitz P: Reducing Opioid Prescriptions by Identifying Responders on Topical Analgesic Treatment Using an Individualized Medicine and Predictive Analytics Approach. J Pain Res 2020;13:1255–1266
67. Soltani M, Farahmand M, Pourghaderi AR: Machine learning-based demand forecasting in cancer palliative care home hospitalization. J Biomed Inform 2022;130:104075
68. Sandham MH, Hedgecock EA, Siegert RJ, Narayanan A, Hocaoglu MB, Higginson IJ: Intelligent Palliative Care Based on Patient-Reported Outcome Measures. J Pain Symptom Manage 2022;63:747–757
69. Wang L, Sha L, Lakin JR, Bynum J, Bates DW, Hong P, et al: Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients With Dementia for Earlier Palliative Care Interventions. JAMA Netw Open 2019;2:e196972
70. Guo W, Gao G, Dai J, Sun Q: Prediction of Lung Infection during Palliative Chemotherapy of Lung Cancer Based on Artificial Neural Network. Comput Math Methods Med 2022;2022:1–7; DOI: 10.1155/2022/4312117
71. Guan H, Dong L, Zhao A: Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making. Behav Sci 2022;12:343
72. Vu E, Steinmann N, Schröder C, Förster R, Aebersold DM, Eychmüller S, et al: Applications of Machine Learning in Palliative Care: A Systematic Review. Cancers 2023;15:1596
73. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J: Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021;47:750–760
74. van de Sande D, van Genderen ME, Smit JM, Huiskens J, Visser JJ, Veen RER, et al: Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ Health Care Inform 2022;29:e100495
75. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG: Right Data Right Now Collaborators: Machine learning in intensive care medicine: ready for take-off? Intensive Care Med 2020;46:1486–1488
76. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al: External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med 2021;181:1065–1070
77. Food and Druck Administration (FDA): Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. https://www.fda.gov/media/145022/download (Zugriffsdatum: 15.11.2023)
78. Food and Druck Administration (FDA): Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) – Discussion Paper and Request for Feedback. 2019. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (Zugriffsdatum: 15.11.2023)
79. Kostick-Quenet KM, Gerke S: AI in the hands of imperfect users. NPJ Digit Med 2022;5:197
80. Friedman B, Kahn PH, Borning A, Huldtgren A: Value Sensitive Design and Information Systems. In: Doorn N, Schuurbiers D, van de Poel I, Gorman ME (Hrsg): Early engagement and new technologies: Opening up the laboratory, Dordrecht: Springer Netherlands 2013;55–95
81. Collins GS, Reitsma JB, Altman DG, Moons KGM: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594
82. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al: PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51–58
83. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, 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
84. Deutscher Ethikrat: Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Berlin: Deutscher Ethikrat 2023. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf (Zugriffsdatum: 01.04.2023)
85. Europäische Kommission: Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. 2021. https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206&from=FR (Zugriffsdatum: 01.04.2023)
86. Du-Harpur X, Watt FM, Luscombe NM, Lynch MD: What is AI? Applications of artificial intelligence to dermatology. Br J Dermatol 2020;183:423–430
87. Bienefeld N, Kolbe M, Camen G, Huser D, Buehler PK: Human-AI teaming: leveraging transactive memory and speaking up for enhanced team effectiveness. Front Psychol 2023;14:1208019
88. Hinck D, Friemert B: Künstliche Intelligenz, Robotik und Digitalisierung im Konzept der Einsatzchirurgie des deutschen Sanitätsdienstes. Chirurg 2020;91:240–244
89. Lee YJ, Cowan A, Tankard A: Peptide Toxins as Biothreats and the Potential for AI Systems to Enhance Biosecurity. Front Bioeng Biotechnol 2022;10:860390
90. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering 2018;2:749–760 https://doi.org/10.1038/s41551-018-0304-0.

Literaturrecherche

Suchhistorie der Pubmed-Literaturrecherche: englische Suchbegriffe „anaesthesia“ (n = 793), „intensive care medicine“ (n = 1650), „emergency medicine“ (n = 2063), „pain therapy“ (n = 680), „palliative care“ (n = 208) zusammen mit „machine learning“ oder „artificial intelligence“. Als Kernthemen wurden im Screening der Literatur das akute Lungenversagen („acute respiratory distress syndrome“ (ARDS),n = 218), Reanimation („resuscitation“, n = 407), „Sepsis“ (n = 877), intraoperative Wachheit („intraoperative awareness“, n = 10), „Polytrauma“ (n = 150), Organversagen („organ failure“, n = 372), Atemwegsmanagement („airway management“, n = 242), Neuralgie („neuralgia“, n = 39), neuropathischer Schmerz („neuropathic pain”, n = 68) und „Herpes Zoster“ (n = 24) gewählt.