Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Sichere KI-Systeme für die Medizin

Datenmanagement und IT-Sicherheit in der Krebsbehandlung der Zukunft

WHITEPAPER

Jörn Müller-Quade et al.
AG IT-Sicherheit,
Privacy, Recht und Ethik

Inhalt

Zusammenfassung

Der Einsatz von Künstlicher Intelligenz (KI) verspricht in der Medizin große Verbesserungen. Lernende Systeme können künftig bei der Prävention, frühzeitigen Diagnose sowie der patientengerechten Therapie zu besseren Behandlungsergebnissen führen und somit unsere Gesundheitsfürsorge verbessern. Durch die Nutzung von patientenindividuellen medizinischen Daten und KI-Assistenzsystemen lassen sich künftig neue medizinische Zusammenhänge entdecken, innovative Präventionsansätze entwickeln, schneller Diagnosen stellen und seltene Erkrankungen früher erkennen. Der Einsatz von KI-Systemen kann zudem Ärztinnen und Ärzte sowie medizinisches Pflegepersonal bei einer verbesserten Versorgung von Patientinnen und Patienten unterstützen und das medizinische Personal entlasten. Zudem ermöglichen KI-basierte medizinische Systeme differenziertere Behandlungsmethoden sowie bessere Ergebnisse in der Vor- und Nachsorge. Konkrete Beispiele für die durch KI-Systeme verbesserten Diagnose- und Therapiemöglichkeiten liefert das fiktive Anwendungsszenario „Mit KI gegen Krebs" (siehe Kapitel 2), das von der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege der Plattform Lernende Systeme entwickelt wurde.

Gleichzeitig stellt der Einsatz von intelligenten und selbstlernenden Systemen im Gesundheitswesen hohe Anforderungen an die IT-Sicherheit der Systeme (siehe Kapitel 3). Ziel des Papiers ist es daher, Risiken des Einsatzes von KI-Systemen im Gesundheitswesen zu identifizieren und mögliche Lösungsvorschläge dafür aufzuzeigen. Nur so kann Vertrauen in die Sicherheit von KI-unterstützten medizinischen Systemen geschaffen werden – was als Voraussetzung für dessen Nutzung gilt. Mögliche Risiken beim Einsatz Lernender Systeme im Gesundheitsbereich sind fehlerhafte oder bewusst verfälschte Trainingsdaten, Angriffe auf die KI-Software, Verletzungen der Privatsphäre der Patientinnen und Patienten sowie Angriffe auf KI-Datenbanken und die fehlende Integration in die klinische Praxis. Entlang des Anwendungsszenarios „Mit KI gegen Krebs" werden technische und organisatorische Bedingungen identifiziert, die für den Einsatz von KI-Assistenzsystemen in der Medizin notwendig sind. Dazu zählen unter anderem die Zertifizierung von KI-Systemen – etwa für die Sicherstellung von unverfälschten Trainingsdaten – sowie besondere Zugriffskontrollmechanismen zum Schutz vor Angriffen auf die KI-Software. Sichere KI-Datenbanken für die KI-basierten Analysen im Gesundheitswesen erfordern auch die Integrität der Datensätze und sichere Übertragungswege. Nach Ansicht der Autoren des vorliegenden Whitepapers sollte die regulatorische Aufsicht der KI-Analyseverfahren samt zugehöriger Trainings- und Testdatensätze an staatlich beauftragte neutrale Einrichtungen vergeben werden. Darüber hinaus sind Lernende Systeme umso präziser und hilfreicher, je mehr Trainingsdaten verwendet werden können. Daher ist es notwendig, medizinische Trainingsdaten aus vielen Studien oder Krankenhäusern unter Schutz der Privatsphäre zu poolen.

Ausgehend von der Analyse des Anwendungsszenarios „Mit KI gegen Krebs“ formulieren die Experten anschließend rechtlich-regulatorische Gestaltungserfordernisse und mögliche Gestaltungsoptionen (siehe Kapitel 4). Dabei liegt der Fokus auf der Frage der Qualitätsabsicherung der für das Training von KI-Systemen verwendeten Daten, der Nachvollziehbarkeit und Erklärbarkeit von KI-Systemen sowie deren Sicherheit im Sinne von Safety und IT-Security. Verknüpft wird dies mit gesellschaftsrelevanten Fragestellungen, etwa zum Nutzen und zu potentiellen Risiken des Einsatzes von KI-Systemen im Gesundheitswesen und zur Verwendung von anonymisierten oder pseudonymisierten Daten. Mit ihrem Whitepaper leisten die Experten der Plattform Lernende Systeme einen Beitrag zur Debatte um sichere KI-Systeme im Gesundheitswesen.

1. Einleitung

Algorithmen helfen heilen

Lernende Systeme versprechen einen großen Nutzen für das komplette Gesundheitswesen. In der Forschung hat es in den vergangenen Jahren zahlreiche Fortschritte gegeben, so beispielsweise in der Onkologie,[1] der Psychiatrie[2] oder der rehabilitativen Robotik.[3] Diese sind auch daran erkennbar, dass die US-amerikanische Food und Drug Administration (FDA) seit dem Jahr 2014 46 Algorithmen-basierte Medizinprodukte zugelassen hat (Stand Juni 2019).[4]

Weitere Fortschritte in der Forschung zeichnen sich bereits ab. In immer mehr Medizingeräten steckt KI. Und trotz der Sorge um die eigenen Daten sind viele Menschen offen dafür, neue Technologien anzuwenden, wenn es um ihre Gesundheit geht. So zeigt etwa eine Studie mit 12.000 Befragten in Europa, dem Nahen Osten und Afrika, dass 54 Prozent der Befragten bereit sind, Robotik und KI in die medizinische Behandlung einzubeziehen (PWC 2017). Lernende Systeme versprechen, die Diagnose und Therapie grundlegend zu verändern und die Prognosegüte für Patientinnen und Patienten deutlich zu verbessern.[5]

Mit Hilfe von KI-Systemen können beispielsweise **große Datenmengen** besser ausgewertet werden. Der Einsatz von KI-Assistenzsystemen verspricht nicht nur, Daten effizienter auswerten, sondern auch neue Ergebnisse generieren zu können. Im Zusammenspiel von KI-Systemen und der ärztlichen Kompetenz können **neue medizinische Zusammenhänge** entdeckt, neue Präventionsansätze entwickelt und seltene Erbkrankheiten erforscht werden. Ärztinnen und Ärzte haben durch KI-Assistenzsysteme Zugang zu exponentiell wachsendem Wissen. Gleichzeitig kann dieses Wissen auch rascher in die breite Versorgung gelangen.

In der **Diagnostik** ermöglicht der Einsatz von Algorithmen eine höhere **Qualität und Zuverlässigkeit**. So können sich Ärztinnen und Ärzte bei ihrer Entscheidung auf zusätzliche patientenindividuelle sowie weltweit vorhandene medizinische Daten stützen und bessere Entscheidungen treffen. Durch den Zugriff auf die elektronische Patientenakte und KI-Assistenzsysteme bei der Diagnose und der Therapie können Ärztinnen und Ärzte schneller die richtigen Diagnosen treffen – somit steigt auch die **Effizienz**. Hinzu kommt, dass KI-Systeme nicht ermüden – sie können rund um die Uhr im Einsatz sein. Mit jedem neuen Trainingsdatensatz wird das KI-System zunehmend wertvoller. Mit Hilfe des Einsatzes von KI-Assistenzsystemen können Ärztinnen und Ärzte **seltene Erkrankungen** früher und eindeutiger erkennen. Frühzeitigere Diagnosen und individuellere Therapien wiederum sorgen für **bessere Behandlungsergebnisse**. Der Einsatz von KI-Assistenzsystemen

1 Für Fortschritte in der Onkologie siehe z. B. Vial et al. 2018.
2 Für Fortschritte in der Psychiatrie siehe z. B. Corcoran et al. 2018.
3 Für Fortschritte in der rehabilitativen Robotik siehe z. B. Kim et al. 2017.
4 Für einen Überblick über die einzelnen Forschungsgebiete siehe The Medical Futurist 2019.
5 Für einen Überblick über die verschiedenen Nutzenversprechen siehe Plattform Lernende Systeme 2019.

ermöglicht differenziertere Untersuchungsmethoden und eine effizientere Auswertung von Vitalparametern – bessere Ergebnisse werden dadurch auch in der Früherkennung sowie der Vor- und Nachsorge erwartet. Wichtig ist aber: Die KI-basierten Systeme nehmen eine unterstützende Rolle ein. Es ist dabei nicht das Ziel, Ärztinnen und Ärzte zu ersetzen.

Notwendige Voraussetzungen für den Einsatz von KI-Systemen im Gesundheitswesen

Lernende Systeme im Gesundheitsbereich versprechen ein sehr großes Potential. Die Anwendung von KI-Systemen in kritischen Anwendungsbereichen[6] ist allerdings an einige Voraussetzungen geknüpft. Denn mit der Technologie gehen auch neuartige, spezifische KI-bezogene Risiken einher, beispielsweise das Training mit nicht-repräsentativen Datensätzen oder die häufig fehlende Nachvollziehbarkeit und Erklärbarkeit von Entscheidungen. Bevor KI-Systeme im Gesundheitsbereich eingesetzt werden, müssen Lösungen entwickelt werden, um diese Risiken zu erkennen und zu beseitigen.[7]

Die genauen Anforderungen an ein KI-System im Gesundheitswesen orientieren sich daran, ob dieses kontinuierlich weiterlernt oder nicht. Ist sein Lernprozess bereits vor der Zulassung abgeschlossen, so besteht die Hauptherausforderung darin, eine Transparenz der Ergebnisse herzustellen. Denn auch wenn das System nachweislich gute Ergebnisse liefert, muss gezeigt werden können, warum ein bestimmtes Ergebnis erzielt wird (Kausalität). Die Nachvollziehbarkeit wird perspektivisch ein wichtiges Kriterium für die Weiterverbreitung und Nutzung des KI-Systems darstellen. Denn Ärztinnen und Ärzte, Pflegerinnen und Pfleger sowie Patientinnen und Patienten wollen wissen, aufgrund welcher Befunde ein Ergebnis erzielt wurde (v. a. auch vor dem Hintergrund möglicherweise fehlerhafter Primärdaten). Dies gilt vor allem für das Deep Learning (DL): Bei dieser KI-Methode fehlen beim heutigen Stand der Technik, im Gegensatz zu Entscheidungsbäumen, die Nachvollziehbarkeit und Erklärbarkeit. Für das medizinische und pflegerische Personal entstehen dadurch haftungsrechtliche Probleme, die es zu lösen gilt.

Lernt das KI-System nach der Zulassung im Einsatz weiter, stellt sich das Problem der Nachvollziehbarkeit von Ergebnissen in verschärfter Form. Das System lernt dann ohne menschliche Überwachung. Die Software nutzt eine kontinuierlich lernende Algorithmik und verarbeitet im Betrieb neue Inputdaten – das KI-System verändert sich also ständig und entspricht daher nicht mehr dem Zulassungsmoment. Zudem schreitet der Erkenntnisstand von Wissenschaft und Technik voran, auch dies beeinflusst das Ergebnis der KI-Software. Eine Verwendung von unüberwacht weiterlernenden, auf KI basierenden Medizinprodukten ist auf der Basis der geltenden Gesetzgebung nicht möglich.[8]

6 Unter „kritischen Anwendungsbereichen“ verstehen die Autoren Anwendungszusammenhänge, die eine Verletzbarkeit des menschlichen Körpers mit sich bringen können.
7 Für einen Überblick über zu klärende Fragen beim Einsatz von KI-Systemen im medizinischen Bereich in den USA siehe Price 2017 sowie die Empfehlungen der Heads of Medicines Agencies und der European Medicines Agency 2019.
8 Grundlage hierfür ist die EU-Medizinprodukteverordnung.

Mit diesem Whitepaper wollen die Autoren einen Beitrag zur Debatte um sichere KI-Systeme im Gesundheitswesen leisten. Sie legen den Fokus auf die Frage der Qualitätsabsicherung der für das Training von KI-Systemen verwendeten Daten, der Nachvollziehbarkeit und Erklärbarkeit von KI-Systemen sowie deren Sicherheit im Sinne von Safety und IT-Security. Hierfür analysieren sie das Anwendungsszenario „Mit KI gegen Krebs" (erarbeitet von der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege der Plattform Lernende Systeme) auf seine Umsetzbarkeit. Ziel ist es, Gestaltungsoptionen für dessen Umsetzung abzuleiten. Hierfür stellen sie in Kapitel 2 das Anwendungsszenario und die Dateninteraktionen vor. Das Kapitel 3 beleuchtet sicherheitsrelevante Aspekte, wie beispielsweise sichere KI-Datenbanken oder unverfälschte Trainingsdaten. Aus der Analyse leiten die Autoren in Kapitel 4 Gestaltungsoptionen und gesellschaftlich zu erörtende offene Fragestellungen ab.

Dieses Papier fokussiert das Datenmanagement und die Sicherheitsaspekte beim Einsatz von KI-Systemen im Gesundheitswesen und widmet sich daher primär technischen Fragen. Die technische Analyse ist ein grundlegender Schritt, um regulatorische Fragen diskutieren und beantworten zu können. KI und Maschinelles Lernen werfen auch gesellschaftsrelevante Fragestellungen auf, die sich nicht rein technisch beantworten lassen.

Ziel des vorliegenden Whitepapers ist es, technische und organisatorische Bedingungen zu identifizieren, die für eine Realisierung des von der Plattform Lernende Systeme entwickelten Anwendungsszenarios „Mit KI gegen Krebs" notwendig sind. Mit einer ersten technisch-organisatorisch fundierten Analyse soll die Grundlage für eine Folgediskussion gelegt werden. Denn die weitere Gestaltung des digitalisierten Gesundheitswesens hat auch Relevanz für das hier zugrunde liegende Anwendungsszenario. Folgende Aspekte wurden daher im Anwendungsszenario offengelassen:

- **Zugriffsberechtigung für Dritte auf die Einträge der elektronischen Patientenakte (ePA):** Momentan ist offen, wie die Patientin oder der Patient Zugriffen auf seine ePA zustimmt. Unklar ist, ob von einem Regel-Ausnahme-Verhältnis zu Gunsten der Nutzung der Daten durch die Ärztin oder den Arzt ausgegangen werden sollte. Es erscheint ebenfalls möglich, dass die Patientin oder der Patient der Nutzung ihrer oder seiner Daten aktiv zustimmen sollte. Sinnvoll könnte auch sein, den Zugriff auf die ePA von Dritten für die Patientin oder den Patienten zu dokumentieren, um Transparenz herzustellen.

- **Freiwillige und geschützte Datenfreigabe:** Die freiwillige und geschützte Datenfreigabe wird derzeit häufig als „Datenspende" diskutiert. Sie ist allerdings noch nicht verbindlich definiert, sodass sich wichtige rechtliche Fragen noch nicht abschließend beantworten lassen.

- **Prüfbedarf der Rechtslage:** Es ist zu prüfen, ob die aktuell bestehenden rechtlichen Vorgaben für den Einsatz von KI-Systemen im Gesundheitswesen ausreichend sind oder ob diese einer Anpassung bedürfen. In Bezug auf Zulassungs- und Rückrufprozesse gilt das für die Europäische Medizinprodukteverordnung (MDR) und das Medizinproduktegesetz (MPG). Zu prüfen ist ebenfalls, ob die aktuellen Datenschutz- und Privacy-Vorschriften im Sinne eines öffentlichen wie patientenindividuellen Interesses anzupassen sind.

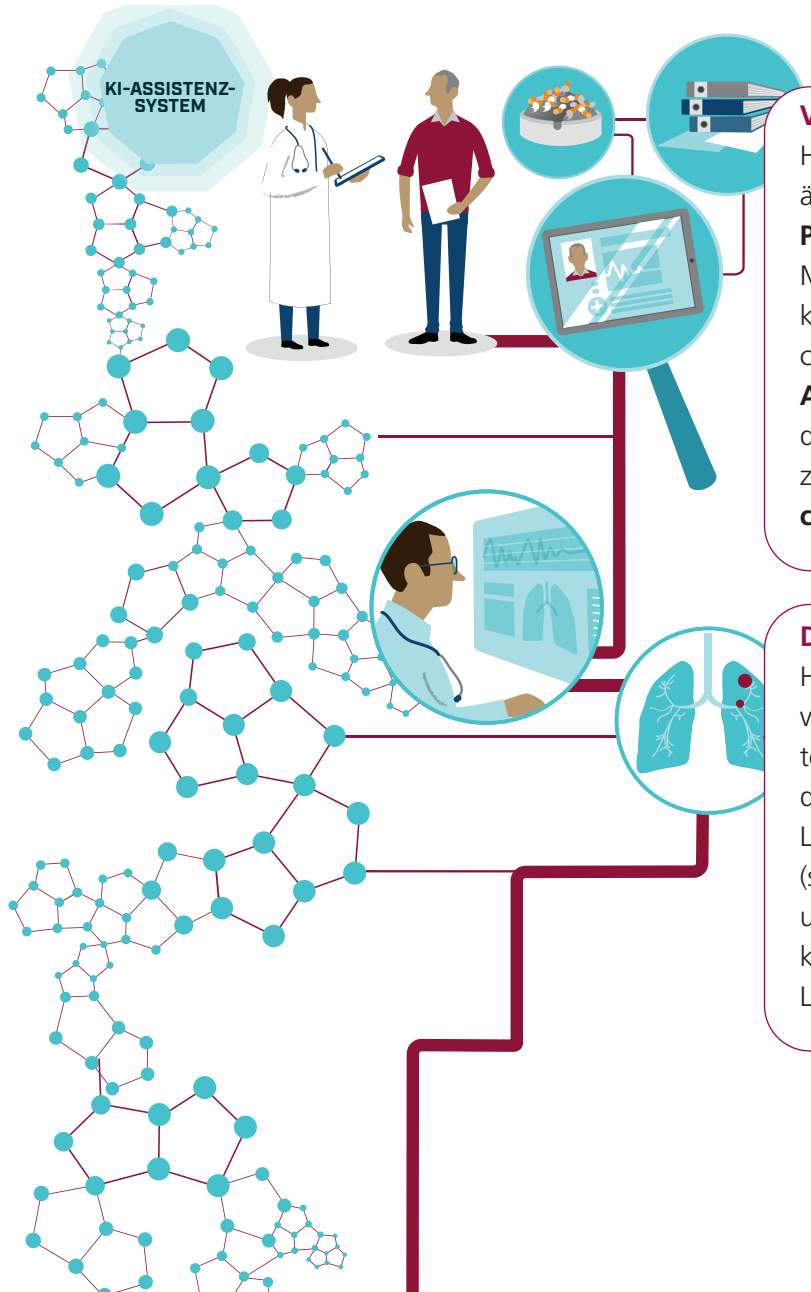
Die angesprochenen Aspekte sind aktuell noch ungeklärt und bedürfen deshalb einer ausführlichen Betrachtung, die im Rahmen vorliegender Publikation nicht geleistet werden kann.

Wie die offenen Aspekte gelöst werden können, ist derzeit noch offen. Die oben skizzierten Lösungsvorschläge stellen hierbei eine Möglichkeit dar, gleichwohl sind andere Lösungsansätze denkbar.

2. KI-Systeme in der Medizin

Das vorliegende Whitepaper basiert auf dem fiktiven Anwendungsszenario „Mit KI gegen Krebs“ der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege der Plattform Lernende Systeme. Das Szenario verspricht verbesserte Diagnose- und Behandlungsmöglichkeiten für Krebspatientinnen und -patienten durch den Einsatz von KI-Systemen.

Im Jahr 2024 erkrankt Anton Merk an Lungenkrebs. Der 65-Jährige ist einer der ersten Patienten, die mit Hilfe eines medizinischen KI-Assistenzsystems behandelt werden. Künftig werden alle behandelnden Ärztinnen und Ärzte darauf zugreifen können – von der Vorsorge über die Diagnose und Therapie bis hin zur Nachsorge. Herr Merk sowie weitere Patientinnen und Patienten werden dadurch eine deutlich verbesserte Überlebens- beziehungsweise Heilungschance haben.

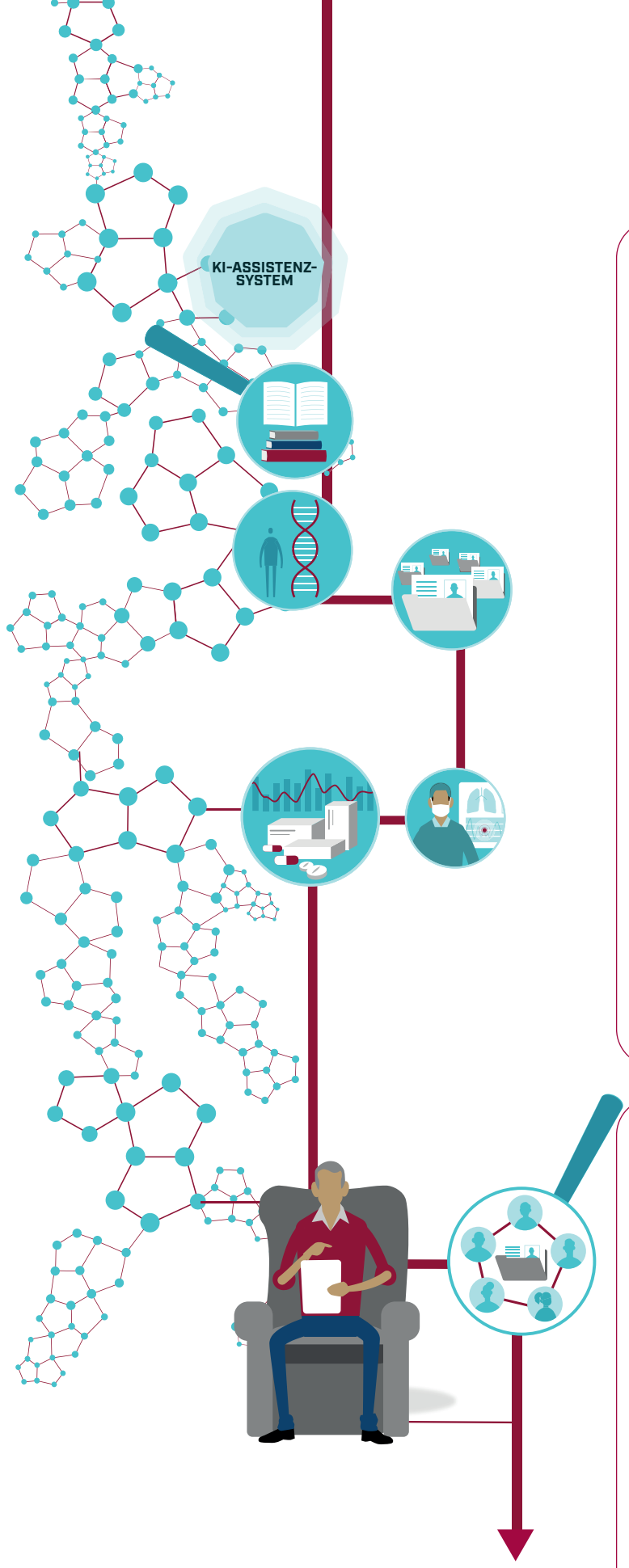


Vorsorge

Herr Merk hat einen Termin bei seiner Hausärztin. Sie hat Einblick in die **elektronische Patientenakte** (siehe 3.5), die auf Herrn Merks Wunsch Informationen zu Vorerkrankungen, aber zum Beispiel auch zu seiner Rauchervergangenheit enthält. Das **KI-basierte Assistenzsystem** (siehe 3.2 und 3.4), mit dem die Hausärztin auf die Patientenakte zugreift, **empfiehlt eine Vorsorgeuntersuchung** (siehe 3.1) bei einem Lungenfacharzt.

Diagnose

Herr Merk stellt sich bei einem Lungenfacharzt vor. Dieser durchleuchtet mit einem Computertomographen (CT) Herrn Merks Lunge. Bei der Auswertung der CT-Bilder wird auch der Lungenfacharzt von dem **KI-Assistenzsystem** (siehe 3.2 und 3.4) unterstützt. Nach dieser und weiteren Untersuchungen in einer Lungenklinik steht die Diagnose fest: Herr Merk hat Lungenkrebs.



Therapie

Das **KI-Assistenzsystem** (siehe 3.2 und 3.4) prüft die vorliegenden Befunde und **empfiehlt** (siehe 3.1), den Tumor chirurgisch entfernen zu lassen. Eine Konferenz von Lungenfachärztinnen und -fachärzten, Strahlentherapeutinnen und -therapeuten sowie Chirurginnen und Chirurgen – das sogenannte Tumorboard – rät Herrn Merk auf dieser Basis zur Operation. Bei der Operation begleitet ein **KI-basiertes Navigationssystem** (siehe 3.2) die Chirurginnen und Chirurgen. Nach dem erfolgreichen Eingriff bespricht Herr Merk die medikamentöse Behandlung mit seinem Lungenfacharzt. Dieser zieht das **KI-Assistenzsystem** (siehe 3.2) heran, das auf umfangreiche Leitlinien, die genetischen Merkmale des Tumors sowie **weltweite Patientendaten** (siehe 3.3) zurückgreift, um den Erfolg unterschiedlicher **Therapie-Alternativen vorherzusagen** (siehe 3.1). Gemeinsam wählen Herr Merk und der Facharzt jene Chemotherapie aus, die für ihn persönlich das beste Verhältnis aus Wirksamkeit und Nebenwirkungen erwarten lässt.

Freiwillige und geschützte Datenfreigabe

Die Behandlung ist optimal verlaufen. Herr Merk möchte, dass die gesammelten Daten der letzten Monate in seine **elektronische Patientenakte** (siehe 3.5) einfließen. Dann sind seine Krankheit und die Behandlung lückenlos dokumentiert und mögliche Auffälligkeiten lassen sich in Zukunft frühzeitig erkennen. Außerdem hat er einer **freiwilligen und geschützten Datenfreigabe** (siehe 3.3) zugestimmt: So stehen seine **Daten anonymisiert und datenschutzkonform** (siehe 3.5) der Forschung zur Verfügung und können weiteren Lungenkrebspatientinnen und -patienten in Zukunft bessere Heilungschancen eröffnen.

Um den Einfluss von KI-Systemen besser einordnen zu können, werden im Folgenden Dateninteraktionen und Rahmenbedingungen im Szenario vorgestellt. Ebenso wird im Kapitel ein Rollenmodell zur Definition von Zuständigkeiten vorgestellt.

2.1 „Mit KI gegen Krebs“: Verbesserte Diagnose und Therapie für Krebspatientinnen und -patienten

Das Anwendungsszenario veranschaulicht den fiktiven Behandlungsablauf eines Lungenkrebspatienten entlang verschiedener diagnostischer und therapeutischer Stationen.

Bei den Untersuchungen werden unterschiedliche fachspezifische KI-Systeme eingesetzt. Sie unterstützen die Medizinerinnen und Mediziner dabei, in diagnostischen Verfahren Krankheitsbilder zu klassifizieren und eine personalisierte Prädiktion des Therapieverlaufes zu bestimmen.

KI-Systeme können sowohl die Effizienz als auch die Qualität der diagnostischen Auswertungen und Abläufe kontinuierlich verbessern. Mit ihrer Hilfe können Krankheiten schneller erkannt werden, denn KI-Systeme können Bilddaten automatisch erheben und strukturieren, daraus Informationen extrahieren und diese mit Vorbefunden abgleichen.

Auch bei der Wahl einer Krebstherapie[9] kann ein KI-System die Ärztinnen und Ärzte dabei unterstützen, die bestmögliche Behandlungsstrategie und die optimale Wirkstoffkombination von Medikamenten zu finden. Darüber hinaus kann das KI-System leitlinienkonforme Empfehlungen entwickeln, indem es zum Beispiel ähnliche Fälle analysiert und auswertet. Zudem ist es möglich, ein digitales Abbild des Patienten in die Behandlung zu integrieren. Dieses Abbild enthält Informationen, die die behandelnden Ärztinnen und Ärzte erhoben haben; anatomische, physiologische, mechanische, pathologische, chirurgische, metabolische, genetische und radiologische Informationen ebenso wie soziologische und psychologische Werte. Diese Komplexität des digitalen Abbildes ermöglicht eine maßgeschneiderte Therapie.

Das Szenario „Mit KI gegen Krebs“ verläuft entlang der sequenziellen Stationen Vorsorge, Diagnose und Therapie. Es endet mit einer freiwilligen und geschützten Datenfreigabe. Sie ermöglicht, dass patientenindividuelle Daten in einer gemeinsamen Forschungsdatenbasis eingepflegt und gespeichert werden. Während der einzelnen Stationen interagiert der Patient mit verschiedenen Leistungserbringern des Gesundheitswesens. Dies sind neben der Hausärztin (Vorsorge) und dem Facharzt (Diagnose) auch die multidisziplinären Expertinnen und Experten im Krankenhaus (Therapie). Bei jeder Station unterstützen verschiedene KI-Assistenzsysteme die Ärztinnen und Ärzte dabei, die Situation zu bewerten und Entscheidungen zu treffen.

9 Für einen Überblick über die verschiedenen Nutzenversprechen siehe den Bericht der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege der Plattform Lernende Systeme 2019.

2.2 Dateninteraktionen im Anwendungsszenario „Mit KI gegen Krebs“

Die KI-Assistenzsysteme sollen effizient, zielgerichtet und mit dem optimalen Ergebnis für die Patientinnen und Patienten zum Einsatz kommen. Damit das gelingt, müssen die Medizinerinnen und Mediziner beziehungsweise Krankenhäuser mit den KI-Algorithmen und den patientenspezifischen Daten auf verschiedene Weise interagieren. Hieraus lassen sich eine Reihe von Dateninteraktionsszenarien ableiten, die für die IT-Sicherheit relevant sind: Dateneingaben, Datenausgaben, Datenspeicherungen, Datenübertragungen und Datenverarbeitungen. Gleichzeitig ergeben sich daraus Fragen zu Aspekten des Datenschutzes und zur Übertragung von Datenzugriffsrechten an Dritte (siehe dazu auch Kapitel 2.3).

Das fiktive Anwendungsszenario basiert auf verschiedenen Daten- und Informationsflüssen, die eine individualisierte Krebsbehandlung ermöglichen. Folgende Dateninteraktionen sind in den einzelnen Stationen relevant.

Vorsorge bei der Hausärztin

- Die Hausärztin ruft die Daten aus der elektronischen Patientenakte (ePA, siehe Infobox Seite 28) ab. Die aktuelle Ärztin und vorhergehende Hausärztinnen und Hausärzte haben die Daten bereits bei früheren Terminen erhoben. Die Daten liegen in elektronisch übertragbarer Form auf einem dauerhaft zugriffsgeschützten Speichermedium bei einem unabhängigen Dritten vor. Die Hausärztin kann zur Unterstützung der Leistungserbringung auf die Daten zugreifen.

- Die Hausärztin gibt neue diagnostische und therapeutische Patientendaten in die ePA ein und speichert sie dort.

Diagnose beim Facharzt

- Daten werden aus der ePA an den Facharzt ausgegeben. Der Facharzt erhält Zugriff auf die ePA des Patienten, also auf die Daten, die die Hausärztin bei der Vorsorgeuntersuchung erhoben hat.

- Das KI-Assistenzsystem verarbeitet die Daten, die der Arzt aufgerufen und erhoben hat. Der Facharzt analysiert den Datenbestand des Patienten mit einer fachspezifischen KI. Die Daten werden zur Auswertung zusätzlicher Diagnostik verarbeitet. Der Facharzt führt mit Hilfe bildgebender Verfahren erweiterte diagnostische Untersuchungen und Analysen durch, um das Krankheitsbild zu spezifizieren. Die CT-Bilder werden mit Hilfe eines weiteren KI-Assistenzsystems analysiert.

Therapie im Krankenhaus

- **KI unterstützt bei der Operation:** Nach der interdisziplinären Entscheidung für die operative Entfernung des Tumors wird das OP-Team durch ein KI-Assistenzsystem bei der patientenspezifisch-minimalinvasiven Zugangsplanung zum OP-Gebiet und bei der intraoperativen Ausführung der chirurgischen Arbeitsschritte unterstützt.

- **Medikamentöse Behandlung:** Für die postoperative Chemotherapie bestimmen die Ärztinnen und Ärzte mit Hilfe der KI die erfolgversprechendste Wirkstoffkombination mit den minimalen Nebenwirkungen. Zur Überwachung des Therapieverlaufs nutzen sie ein weiteres Assistenzsystem.

Fachspezifische KI-Systeme
In dem beschriebenen Anwendungsfall werden vier verschiedene fachspezifische KI-Systeme betrachtet:
- Allgemeiner Vorsorgedaten-Assistent zur Bewertung der ePA-Daten
- CT-Bild-Auswerte-Assistent
- Operations-Assistent für die chirurgische Krebsbehandlung
- Berechnungs-Assistent für die Wirkstoffkombination der Chemotherapie

Weitere Datennutzung

- **Daten zusammenführen:** Der Patient stimmt einer Zusammenführung seiner verfügbaren Gesundheitsdaten zu. Sein Behandlungsablauf ist dadurch lückenlos dokumentiert und kann gegebenenfalls weiteren Fachärztinnen und Fachärzten zur Verfügung gestellt werden.

- **Freiwillige und geschützte Datenfreigabe:** Der Patient entscheidet sich dafür, seine Daten anonymisiert (Definition siehe Infobox Seite 14) für die Forschung oder das Training des KI-Systems zur Verfügung zu stellen. Eine Alternative zur anonymisierten Bereitstellung der Daten ist die Pseudonymisierung (Definition siehe Infobox Seite 14). Hierbei können die gewonnenen Daten zu einem späteren Zeitpunkt wieder zusammengeführt werden. Zentrale Fragen lauten: Wer verfügt bei einer Pseudonymisierung über den Schlüssel und ist berechtigt, bei Bedarf und vorliegender Zustimmung der Patientin oder des Patienten die Daten nachträglich wieder zusammenzuführen?

Anonymisierung und Pseudonymisierung von Daten
Bei der Anonymisierung werden personenbezogene Daten derart verändert, dass deren Zuordnung zu einer Informationsgeberin oder einem Informationsgeber nicht mehr möglich ist. So ist beispielsweise bei einer geheimen Wahl keine Zuordnung des Wahlzettels zur Wählerin oder zum Wähler möglich, da keine Referenz gegeben ist. Bei der Pseudonymisierung werden Identifikationsmerkmale der Person nicht vernichtet, sondern durch ein Pseudonym (meist ein Code oder ein Alias) ersetzt. Auch dies erschwert die Zuordnung der Daten zu einer Person, macht sie aber nicht unmöglich. Somit können mehrere Datensätze mit dem gleichen Pseudonym in Zusammenhang gebracht werden, ohne die Identität der Person preiszugeben. Gleichzeitig ist es bei Bedarf, zum Beispiel bei entsprechender medizinischer Indikation, möglich, die Daten wieder einer Person zuzuordnen, weil es einen Schlüssel als Referenz zwischen den getrennt aufbewahrten Daten und Identifikationsmerkmalen gibt. Bei großen pseudonymisierten Datenmengen, die wesentliche Aspekte des Lebens einer Person abdecken (z. B. Krankheitsgeschichte, Einkommen, Wohnort), ist eine Identitätsfeststellung ohne Schlüssel zwar verboten, unter Umständen technisch jedoch möglich. Auch bei anonymisierten Daten ist bei entsprechendem Kontextwissen eine Zurückverfolgbarkeit der Daten theoretisch möglich; dies ist aber deutlich schwieriger als bei pseudonymisierten Daten.

2.3 Rahmenbedingungen

Das Anwendungsszenario zeigt, wie das Zusammenspiel von Mensch, Computer und Maschine im Jahr 2024 die Krebsbehandlung verbessern könnte. Es skizziert die technischen Möglichkeiten, die sich heute schon in der Medizin abzeichnen: Die Technologien, die das Szenario ermöglichen, werden stetig weiterentwickelt; ihre Leistungsfähigkeit schreitet rasant voran. Auf lokalen Datenbeständen sind KI-Algorithmen bereits erfolgreich eingesetzt worden, um Diagnose und Therapie zu unterstützen.[10]

Ob das Anwendungsszenario in naher Zukunft in den medizinischen Alltag gelangt, hängt jedoch nicht ausschließlich von technischen Fragen ab. Gleichzeitig sind organisatorisch-administrative (und unter Umständen legislativ-regulatorische) Rahmenbedingungen erforderlich, die derzeit noch nicht in vollem Umfang gegeben sind.

Ungeklärt sind etwa die folgenden zentralen Fragen:

1. Wo und in welcher Form werden die elektronischen Patientendaten und die Meta-Daten der KI-bewerteten ePA (zwischen-)gespeichert, übertragen und erweitert? Wer stellt die dafür notwendige Infrastruktur bereit?
2. Wie autorisieren die Patientinnen und Patienten den Zugriff auf ihre Daten und deren weitere pseudonymisierte Verwendung, etwa für Forschungszwecke?

10 Für Fortschritte zum Beispiel im Bereich der Kindermedizin siehe z. B. Ärzteblatt GmbH 2019, für Fortschritte im Bereich der Anästhesie siehe z. B. Neumuth/Franke 2018.

3. Wie kann der Einsatz der Technologien auf verteilten Datenbeständen gestaltet werden, auf die Vertreterinnen und Vertreter der verschiedenen am Behandlungsprozess beteiligten Institutionen Zugriff haben müssen?
4. Wie soll der Zulassungsprozess für KI-Assistenzsysteme im Detail gestaltet werden? Wie werden die Rechte und Pflichten zwischen den Beteiligten verteilt?
5. Wer pflegt und trainiert das KI-System und stellt die neueste KI-Software auf Anfrage einer Ärztin oder eines Arztes bereit?

2.4 Rollenmodell für klare Verantwortlichkeiten

Berechtigungen ordnungsgemäß zu vergeben ist der erste elementare Schritt zum Schutz von Systemen. Nur so können IT-Systeme vor unrechtmäßigem Zugriff geschützt und die Qualität und Verfügbarkeit der Daten gewährleistet werden. Das folgende Rollenmodell beschreibt die Verantwortlichkeiten, die die verschiedenen Akteure im Anwendungsszenario „Mit KI gegen Krebs" erfüllen müssen. Es zeigt, welche Rechte und Pflichten sie im fiktiven Anwendungsszenario im Jahr 2024 innehaben, sodass eine sichere medizinische Behandlung garantiert ist.

Akteur	Datenumgang/Rechte	Daten- und Informationsarten
Technischer Betreiber eines KI-Assistenzsystems	• Zentrale Speicherung über Rechenzentrum • Ggf. Korrekturen an Datensätzen nur auf Veranlassung der Hausärztin oder des Hausarztes oder bei erkannten Fehlinformationen des KI-Assistenzsystems • Technische Datenverarbeitung • Partizipation in Prozessen zur ständigen Aktualisierung des KI-Assistenzsystems	Sämtliche Datensätze (zentraler Betrieb) oder nur Aggregat im Sinne von Trainingsdaten (dezentraler Ansatz)
Unabhängiger autorisierter Betreiber des KI-Assistenzsystems	• Staatlich beauftragte neutrale Einrichtung • Verwaltung und Pflege der Analyseverfahren sowie der Datensätze (für genauere Informationen siehe „Sichere KI-Datenbanken", Seite 24 ff). • Keine eigenen Einspeise- oder Veränderungsmöglichkeiten, da ggf. ein eigenes ökonomisches Interesse vorliegen könnte	Sämtliche Datensätze der ePA (sofern kein Einspruch der Patientin oder des Patienten)

Behandelnde Hausärztin und behandelnder Hausarzt	• Rolle als Intermediär zwischen „System"/ Behandlung sowie Patientin oder Patient
	• Dateneingabe via ePA der Patientin oder des Patienten: Erweiterung des Datenbestands der Patientin oder des Patienten durch diagnostische und therapeutische Daten (inkl. KI-basierte Erkenntnisse/Empfehlungen)
	  ◦ Schritt 1: Abruf der ePA-Daten
	  ◦ Schritt 2: Abruf der KI-Software
	  ◦ Schritt 3: Anwendung der KI-Software auf die ePA-Daten
	• Datenausgabe sämtlicher Daten sowie KI-basierte Diagnoseunterstützung
		Sämtliche Datensätze der ePA, Zugriff auf die KI-Software (Patientenvorbehalt)

Akteur	Datenumgang/Rechte	Daten- und Informationsarten
Behandelnde Fachärztin und behandelnder Facharzt	• Datenausgabe (aus der ePA oder direkte Übermittlung KI-basierter Erkenntnisse) an die Fachärztin oder den Facharzt • Die Fachärztin oder der Facharzt erhält spezifischen Zugang (Leserecht) zu erhobenen Daten durch Zugriff auf die ePA der Patientin oder des Patienten • Initiierung einer Datenverarbeitung durch KI-Assistenzsysteme zur Analyse durch die Fachärztin oder den Facharzt • Initiierung Datenverarbeitung durch KI-Assistenzsystem zur zusätzlichen Diagnostik • Dateneingabe via ePA der Patientin oder des Patienten mit zusätzlichen Erkenntnissen und Therapieempfehlungen	Spezifische, auf die Diagnose/Symptome abgestellte Datensätze, Zugriff auf die KI-Software
Sonstige Ärztinnen und Ärzte	• Spezifische Datenausgabe und -eingabe analog zur Fachärztin oder zum Facharzt (z. B. im Rahmen einer Tumorkonferenz)	Bereichsspezifisch notwendige Datenarten und -sätze
Patientin und Patient	• Grundsätzlich sämtliche Datenausgabe- und Leserechte • Individuelle Rechtesteuerung gegenüber Dritten möglich und ggf. auch notwendig (Patientensouveränität) • Freiwillige und geschützte Datenfreigabe • Dateneingabe z. B. von gesundheitsrelevanten Verhaltensdaten	
Krankenhaus allg./andere Bereiche	• Spezifische Datenausgabe analog zur Fachärztin oder zum Facharzt (z. B. zur OP-Unterstützung, postoperative Chemotherapie) • Spezifische Dateneingabe analog zur Fachärztin oder zum Facharzt	Bereichsspezifisch notwendige Datenarten und -sätze z. B. zur OP-Unterstützung, postoperative Chemotherapie oder Erkenntnisse des Therapieverlaufs
Krankenkasse	• Anbieten der ePA • Beauftragung einer dritten Organisation mit der Speicherung, Aktualisierung und Verwaltung der Daten	Krankenkassen haben keinen Zugriff auf die Datensätze der ePA, sie dürfen wegen ihrer ökonomischen Interessen nur Zugriff auf Daten haben, die für die Abrechnung erforderlich sind
Pharmaunternehmen	• Datenausgabe für Forschungszwecke • Dateneingabe von produktspezifischen Erkenntnissen (Schutz vor manipulativer Einspeisung durch eine unabhängige Qualitätsprüfung)	Lediglich Ausgabe aggregierter Daten für Forschung und Entwicklung (F&E-Tätigkeiten)
Akademische Forschungseinrichtungen	• Datenausgabe für Forschungszwecke	Lediglich Ausgabe aggregierter Daten für F&E-Tätigkeiten

3. Anforderungen an IT-Sicherheit im Szenario „Mit KI gegen Krebs“

Wie bereits dargestellt, birgt der Einsatz von KI-Assistenzsystemen im medizinischen Bereich ein großes Nutzenpotential. Gleichzeitig geht der Einsatz dieser Systeme auch mit potentiellen IT-Sicherheitsproblemen einher. So lassen sich beispielsweise medizinische Bilder oder auch Computertomographie-Scans so manipulieren, dass der Algorithmus zu falschen Ergebnissen kommt (Finlayson et al. 2018, Mirsky et al. 2019). Hinzu kommt, dass Dateninputs ebenfalls manipulierbar sind (BSI/ANSSI 2019). Wichtig ist deshalb, dass die Trainingsdaten unverfälscht sind, dass das KI-System/die KI-Datenbank sicher ist und dass auch die elektronische Patientenakte ausreichend geschützt ist. Im folgenden Kapitel werden diese Aspekte diskutiert und weitere Fragen erörtert, wie beispielsweise das Pooling von Daten.

3.1 Originale, unverfälschte Trainingsdaten sicherstellen

Qualitativ hochwertige Trainingsdaten bilden die Grundlage für ein fehlerfrei arbeitendes KI-System. Trainingsdaten können jedoch sowohl unbewusst verzerrt als auch gezielt manipuliert werden. Dieses Kapitel analysiert beide Varianten der Datenverfälschung. Zudem enthält es Vorschläge, wie das Risiko von Verzerrungen und Manipulationen verringert werden kann.

(Unbeabsichtigte) Verzerrungen

Datensätze können einen Bias (Definition siehe Infobox Seite 18) enthalten und somit nicht repräsentativ sein. So werden beispielsweise in den USA Algorithmen eingesetzt, um die Rückfallwahrscheinlichkeit von Angeklagten zu bestimmen. Der Algorithmus wurde allerdings vor allem mit historischen Daten (z. B. aus Kriminalitätsstatistiken) trainiert, die nicht auf kausalen Zusammenhängen, sondern auf statistischen Korrelationen beruhen. Menschen aus bestimmten Bevölkerungsgruppen, die in der Vergangenheit häufiger ins Visier der Strafverfolgungsbehörden geraten waren, erhielten automatisch schlechtere Prognosen.[11] Betroffen waren etwa Angehörige ethnischer Minderheiten oder einkommensschwacher Gruppen. Ein unbewusster Bias bei der Auswahl der Trainingsdaten kann auch dazu führen, dass der Algorithmus Erkrankungen nicht berücksichtigt, die bei bestimmten ethnischen Gruppen häufiger auftreten. So warnt etwa die HMA-EMA-Studie, dass KI-Systeme, die auf Trainingsdaten von kaukasischen Patientinnen und Patienten erhoben wurden, nicht für Betroffene anderer ethnischer Gruppierungen verwendet werden können.[12] Dies ist der Fall, wenn sie vor allem mit Daten anderer ethnischer Gruppen trainiert worden sind.

11 Für einen Überblick über diese Praxis siehe Angwin et al. 2016.
12 Für weitere Informationen siehe Heads of Medicines Agencies und European Medicines Agency 2019, Seite 25.

Bias
Ein Bias bezeichnet allgemein Verzerrungseffekte. Die Psychologie versteht darunter Einstellungen oder Stereotypen, welche die Wahrnehmung unserer Umwelt, Entscheidungen und Handlungen positiv oder negativ beeinflussen. Diese Beeinflussung kann unbewusst (impliziter Bias) oder bewusst (expliziter Bias) geschehen. In der Statistik wird ein Bias als Fehler im Rahmen der Datenerhebung und -verarbeitung (z. B. Fehler in der Stichprobenauswahl) oder die bewusste oder unbewusste Beeinflussung von Probandinnen und Probanden verstanden. Für einen Überblick über unterschiedliche Arten eines Bias siehe Beck et al. 2019.

Gezielte Manipulationen

Datensätze können bewusst manipuliert oder verfälscht werden. Für den Fall gezielter Manipulationen legen die Autoren folgende Prämisse zugrunde (siehe 2.4): Ausschließlich *Leistungserbringer* (Ärztinnen und Ärzte, Kliniken, Krankenhäuser) und *Leistungsträger* (Krankenkassen) sowie technische und unabhängige autorisierte *Betreiber* des KI-Assistenzsystems haben einen Zugang zum KI-System. Nur diese berechtigten Akteure können neue Trainingsdaten erzeugen und an den KI-Systembetreiber weiterleiten. Ebenso können nur sie das KI-System herunterladen, zum Beispiel zum Zwecke einer Patientendiagnoseauswertung. Pharma-Hersteller sowie Patientinnen und Patienten haben hingegen keinen Zugang.

Akteure, die versuchen, Trainingsdaten für KI-Systeme gezielt zu verfälschen oder zu manipulieren, können unterschiedliche **Ziele** verfolgen. So können die KI-Trainingsdaten zum wirtschaftlichen Vorteil für Leistungserbringer und -träger oder für Pharma- und Medizingerätehersteller manipuliert werden; etwa indem das teuerste Medikament die höchste Weiterempfehlungsrate durch das KI-System aufweisen soll. Ziel kann es aber auch sein, das KI-System in Einklang mit Studienergebnissen zu bringen und so besonders hohes Ansehen für eine Gruppe von Forscherinnen und Forschern zu generieren. Auch zerstörerische Absichten von Einzelpersonen können eine Motivation sein.

Die (unbeabsichtigte) **Verzerrung oder** (bewusste) **Manipulation** kann an unterschiedlichen Stellen auftreten – sowohl physisch betrachtet als auch auf den Prozess bezogen. So ist es beispielsweise möglich, dass der KI-Hersteller bereits verzerrte Trainingsdaten erhält und einsetzt. Dies kann unbewusst, etwa aufgrund eines Bias, oder bewusst, zum Beispiel durch vorheriges Verfälschen oder Filtern, geschehen. Es ist denkbar, dass eine Mitarbeiterin oder ein Mitarbeiter des Betreibers gefälschte Trainingsdaten eingibt und so die Ergebnisse der KI-Systeme manipuliert. Verzerrte oder manipulierte Trainingsdaten können auch dadurch entstehen, dass der Leistungserbringer bestimmte Daten bewusst zurückhält. Hinzu kommt die Möglichkeit der Angriffe von außen (siehe 3.2).

Folgen von Verzerrungen und Manipulationen

Die Folge der Manipulation: Das KI-System gibt nach der Trainingsphase auf Grundlage eines fehlerhaften Modells eine falsche und möglicherweise auch problematische Empfehlung aus. Dies könnte in allen Versorgungsstufen des betrachteten Szenarios geschehen.

Bei der Vorsorge könnten verzerrte Trainingsdaten möglicherweise dazu führen, dass das KI-System zu kurze Vorsorgeuntersuchungszyklen vorschlägt. Darüber hinaus könnte fälschlicherweise eine seltene Krankheit vermutet werden, bei der die medizinische Behandlung sehr kostenintensiv ist.

Bei der Diagnose könnte das KI-System auf Basis verzerrter Trainingsdaten falsche Diagnosen oder vor allem Diagnoseverfahren empfehlen, die mit besonders hohen Kosten verbunden sind. Alternativ könnte das KI-System auch (überflüssige) weitere Diagnoseuntersuchungen vorschlagen, die eine andere Fachärztin oder ein anderer Facharzt durchführen müsste.

Bei der tatsächlichen Behandlung (inklusive einer möglichen Operation) könnte das KI-System gegebenenfalls eine OP-Strategie vorschlagen, die mit Blick auf das Krankheitsbild falsch gewählt ist und deshalb die Gesundheit der Patientin oder des Patienten gefährden kann oder unverhältnismäßig hohe Zeit- und Personalressourcen erfordert. Darüber hinaus könnte es gegebenenfalls auch vorkommen, dass das KI-System eine Kombination aus medikamentöser Behandlung und Operation vorschlägt – auch, wenn diese nur unwesentlich bessere Erfolge verspricht. Dies würde ebenfalls zu höheren Kosten führen.

Bei der medizinischen Versorgung könnte eine fehlerhaft trainierte KI-Software ein besonders teures Medikament bestimmter Hersteller empfehlen – auch wenn dieses nicht erfolgsversprechender ist als die preisgünstigeren Alternativen. Ebenso könnte die KI-Software in Folge verzerrter Daten eine zu große Wirkstoffmenge eines bestimmten Präparates empfehlen.

Mögliche Maßnahmen gegen unbeabsichtigte Verzerrungen[13]

- **Vollständigkeitsüberprüfung:** Die Anzahl der Erkrankungen und der jeweiligen Krankheitsbilder sind für einen Betrachtungszeitraum statistisch bekannt. Die Anzahl der neuen Dateneinträge in die KI sollte, zeitlich versetzt, mit der statistisch erfassten Anzahl der Erkrankungen weitestgehend übereinstimmen. Damit kann ermittelt werden, ob einzelne Datenquellen keine neuen Daten geliefert haben und dadurch ein oder mehrere Bias wahrscheinlich werden. Grundsätzlich sollte das KI-System regelmäßig mit Hilfe statistischer Methoden auf mögliche Verzerrungen hin überprüft werden.

13 Die vorgestellten Maßnahmen gegen unbeabsichtigte Verzerrungen und gezielte Manipulationen beziehen sich nur auf überwachtes Lernen, da unüberwachtes Lernen im Medizinbereich aktuell nicht zugelassen ist (siehe EU-Medizinprodukteverordnung).

- **Anwendbarkeitsüberprüfung:** Vor dem Einsatz des KI-Systems muss überprüft werden, ob die spezifischen Daten der Patientin oder des Patienten zu den für das KI-System verwendeten Klassen von Patientinnen und Patienten passen.

Mögliche Maßnahmen gegen gezielte Manipulationen

- **Die Ärztin oder der Arzt als (Letzt-)Entscheider:** Das KI-System muss der behandelnden Ärztin oder dem behandelnden Arzt auf Nachfrage nachvollziehbare Gründe, Entscheidungswege und Wahrscheinlichkeiten präsentieren. Medizinerinnen und Mediziner dürfen die Empfehlung des KI-Systems nicht als Entscheidung verstehen, sondern vielmehr als eine Meinung unter mehreren. Wichtig ist zudem der Austausch mit Kolleginnen und Kollegen. Endgültige Entscheidungen über Diagnosen und Therapieempfehlungen müssen immer die behandelnden Medizinerinnen und Mediziner treffen.

- **Gemeinsame Leitlinien erstellen (wissenschaftlich, medizinisch, ethisch, gesellschaftlich und wirtschaftlich) sowie Empfehlungen des KI-Systems überprüfen:** Eine unabhängige Gruppe, bestehend aus Expertinnen und Experten, sollte zunächst jährlich die von dem KI-System gegebenen Empfehlungen stichprobenartig überprüfen. Stellt sie nach zwei Überprüfungen keine Auffälligkeiten fest, so wird der Zeitraum bis zur nächsten verpflichtenden Überprüfung ausgedehnt.

- **Originalität der Trainingsdaten nachweisen:** Jede berechtigte Datenquelle, wie Krankenhäuser, Krankenkassen, Kliniken etc., muss das Vorgehen bei der Auswahl der Trainingsdaten klar dokumentieren. Die Daten müssen repräsentativ sein und unabhängig geprüft werden können. Diese Prüfung kann in Form einer Kreuzvalidierung auf Daten durchgeführt werden, die die Zielpopulation umfassend repräsentieren. Integrität und Authentizität der Trainingsdatensätze sollen sichergestellt sein.

- **Berechtigung nachweisen:** Nur autorisierte Personen dürfen neue Trainingsdaten in das KI-System hochladen. Die Daten müssen den Originalitätsnachweis besitzen; auch hier müssen die Integrität und Authentizität sichergestellt sein. Die digitale Identität der berechtigten Person stellt eine Schlüsselrolle dar. Ein denkbares sicheres Identifizierungsverfahren wäre die Nutzung des Heilberufsausweises (HBA) und der zugehörigen PIN durch die Ärztin oder den Arzt.

- **KI-System zertifizieren:** Neben der technischen Beschreibung der KI-Software ist eine Prüfvorschrift wichtig. Sie beschreibt, wie die KI-Systeme geprüft werden können. Die KI-Software muss die Prüfung bestehen, um ein Zertifikat zu erhalten. Es ist eine turnusmäßige Überprüfung des Zertifikats vorgesehen, im Sinne eines Re-Assessments. Zudem existiert eine Kennzeichnungspflicht der eingesetzten KI-Systeme, die über Eigenschaften und Zulassung für bestimmte Einsatzgebiete informiert.

3.2 KI-Software vor Angriffen schützen

Wie der vorherige Abschnitt gezeigt hat, ist die Garantie unverfälschter Trainingsdaten ein wesentlicher Aspekt der Safety und IT-Sicherheit. Darüber hinaus ist es wichtig, mögliche Risiken beim Trainingsprozess des Lernenden Systems zu betrachten und Schutzmechanismen für die KI-Software zu analysieren. Ein KI-System mit kontinuierlich lernenden Algorithmen verarbeitet neue Inputdaten mit dem Ziel, das KI-Modell über die Zeit zu verbessern und dessen Präzision zu erhöhen. Es kann hierbei zwischen überwachtem und unüberwachtem Lernen unterschieden werden (Definition siehe Infobox). Lernende Systeme, die auf unüberwachtem Lernen basieren, müssen im Gesundheitswesen zum aktuellen Zeitpunkt grundsätzlich ausgeschlossen werden, da die Ergebnisse vorher unbekannt sind und kaum kontrolliert werden können. Aber auch beim überwachten Lernen wird das System regelmäßig neu trainiert – dies schafft ebenfalls Potential für Manipulationen. Auch bei der Nutzung des KI-Systems bestehen diverse Anreize und Motivationen zur Manipulation, die im vorherigen Abschnitt diskutiert wurden. Zusammenfassend lässt sich sagen: Mit dem Einsatz eines KI-Systems entstehen neue spezifische Angriffsmöglichkeiten. Die relevantesten werden im Folgenden betrachtet.

Überwachtes vs. unüberwachtes Lernen
Beim überwachten Lernen werden Lernalgorithmen verwendet, die als Trainingsmaterial neben Rohdaten auch die erwarteten Ergebnisse erhalten. Weicht die Ausgabe des trainierten Modells vom gewünschten Ergebnis ab – wenn beispielsweise eine Tulpe als Rose identifiziert wird –, passt der Lernalgorithmus das Modell an. Ziel ist es, dem Netz durch unterschiedliche Ein- und Ausgaben die Fähigkeit anzutrainieren, selbst Verbindungen herzustellen.

Beim unüberwachten Lernen werden hingegen Lernalgorithmen verwendet, die kein Prognoseziel, sondern nur die Rohdaten erhalten. Sie erzeugen ein Modell, das die Eingaben abstrakt beschreibt und Vorhersagen ermöglicht. Das Netz erstellt dann selbstständig Klassifikatoren, nach denen es die Eingabemuster einteilt. Ziel ist es, in einem Datensatz interessante und relevante Muster zu erkennen oder die Daten kompakter zu repräsentieren.

Ein Lernendes System, das auf unüberwachtem Lernen basiert, verändert sich über die Zeit: Es lernt kontinuierlich, sodass sich auch das zugrunde liegende Modell über die Zeit wandelt und andere Entscheidungen treffen kann. Dieser sogenannte *Concept Drift* ist ein bekanntes Problem im Bereich des Maschinellen Lernens. Manipulationen, die die unabhängige Qualitätssicherung unter Umständen umgehen, können den *Concept Drift* gezielt beeinflussen. So werden mittels einer manipulierten freiwilligen und geschützten Datenfreigabe neue Trainingsdaten zur Verfügung gestellt, die das Modell im kontinuierlichen Lernprozess gezielt verändern. Möglich wäre es auch, dass Daten vor dem Neu-Trainieren des Systems manipuliert oder Parameter des KI-Modells gezielt verzerrt werden. Deshalb müssen Mechanismen eingesetzt werden, um solche manipulierten Daten zu verhindern

oder frühzeitig erkennen zu können. Der *Concept Drift* des KI-Systems muss überwacht werden – dazu können die in Kapitel 3.1 vorgestellten Methoden zur Erkennung von gezielten Verzerrungen genutzt werden.

KI-Systeme sind anfällig für Inversionsangriffe auf das Modell (engl. *Model Inversion Attack*). Bei diesen Angriffen rekonstruiert die Angreiferin oder der Angreifer einen Teil der Trainingsdaten. Im Rahmen von *Membership Inference Attacks* kann die Angreiferin oder der Angreifer Informationen über die Zugehörigkeit von bestimmten Daten zum Modell erlangen und so feststellen, ob bestimmte Daten während der Trainingsphase des KI-Systems genutzt wurden. Eine Anonymisierung aller genutzten Daten kann das Risiko solcher Angriffe minimieren. Allerdings sollten zusätzliche Maßnahmen getroffen werden, um solche Angriffe zu erschweren oder frühzeitig erkennen zu können.

KI-Systeme können während der Nutzung auch durch sogenannte *Model Stealing Angriffe* attackiert werden: Dabei versucht eine Angreiferin oder ein Angreifer, die Modellparameter des KI-Systems zu stehlen, um an interne Informationen über das KI-Modell zu gelangen. Mit gezielten Anfragen an das KI-System sollen dessen Entscheidungsgrenzen – und somit auch dessen Parameter – erlernt werden. *Model Stealing Angriffe* verfolgen zwei Zielsetzungen: Erstens zielen sie darauf ab, andere Angriffe vorzubereiten, insbesondere zur Manipulation des KI-Systems während der Laufzeit. Zweitens bringt die Angreiferin oder der Angreifer geistiges Eigentum des Betreibers in den eigenen Besitz. Derzeit existieren noch keine Methoden, mit denen KI-Systeme wirksam gegen solche Angriffe geschützt werden können. Mittels einer Analyse der Anfragen an das KI-System können mögliche Angriffe aber eventuell erkannt werden.

Bei einem erfolgreichen *Model Stealing Angriff* erstellt eine Hackerin oder ein Hacker möglicherweise eine Kopie der KI-Software, die dann in anderen Kontexten eingesetzt werden kann. Anwenderinnen und Anwender können möglicherweise nicht erkennen, ob sie mit dem Originalsystem oder einer sehr guten Kopie interagieren. Schutzmechanismen wie etwa besondere Zugriffskontrollmechanismen können das Risiko für solche Angriffe minimieren.

3.3 Trainingsdaten unter Wahrung der Privatsphäre poolen

Lernende Systeme sind umso präziser und hilfreicher, je mehr Trainingsdaten zur Verfügung stehen und verwendet werden können. Seltene medizinische Fälle lassen sich nicht erlernen, wenn sie in den Trainingsdaten nur einzeln auftreten. Ein „Underfitting“ (siehe Infobox Seite 23) der KI-Software sollte daher unbedingt vermieden werden. Ideal wäre es, wenn medizinische Daten aus vielen Studien oder Krankenhäusern gepoolt werden könnten. KI-Systeme könnten dann auf Basis des großen Daten-Pools viel besser Muster lernen und erkennen. Jedoch verbieten zum aktuellen Zeitpunkt Datenschutzvorschriften dieses Vorgehen. Es existieren Techniken, die es erlauben, die Daten (virtuell) zu poolen, ohne dabei die Privatsphäre von Patientinnen und Patienten zu verletzen. Der folgende

Abschnitt erläutert diese Techniken. Drei Herausforderungen bleiben aber weiterhin bestehen: Die Techniken müssen rechtlich anerkannt sein, sie müssen im Einklang mit dem Datenschutzrecht angewendet werden und sie müssen als Gesamtsystem sicher sein.

Underfitting eines Modells
Liegen zu wenige Trainingsdaten vor, kann es passieren, dass das KI-System gegebenenfalls zu wenige Muster aus den Trainingsdaten erlernen kann. In Extremfällen kann das KI-System dann nicht einmal den zugrundeliegenden Trend erkennen. In der Folge weist das Modell nur eine geringe Generalisierbarkeit und Vorhersagekraft auf.

Sichere Mehrparteienberechnungen

Sichere Mehrparteienberechnungen existierten lange Zeit nur als eine theoretische Möglichkeit. Sie sind Berechnungsverfahren, bei denen mehrere Parteien verteilt auf geheimen Eingaben der Parteien ein Ergebnis berechnen können, ohne dass über die geheimen Eingaben mehr bekannt wird, als das Ergebnis ohnehin verrät. Darüber hinaus ist es nicht möglich, das Ergebnis zu manipulieren. Sichere Mehrparteienberechnungen emulieren also eine zentrale vertrauenswürdige Instanz durch verteilte Berechnungen. Inzwischen gibt es aber erstaunlich performante Realisierungen und erste Start-up-Firmen, die sichere Mehrparteienberechnungen anbieten. Trotzdem wird sich erst noch zeigen müssen, für welche Art von Anwendungen die Verfahren in der Praxis effizient genug sind. Sichere Mehrparteienberechnungen sind eine vielversprechende Zukunftstechnologie, aktuell ist der Aufwand für die meisten Anwendungen aber immer noch zu hoch.

Voll-homomorphe Verschlüsselung

Eine voll-homomorphe Verschlüsselung erlaubt ein Rechnen mit verschlüsselten Daten, ohne dass diese dafür entschlüsselt werden müssen. Der verarbeitende Server führt also einen Algorithmus aus, ohne die verwendeten Daten zu erlernen. Mit der voll-homomorphen Verschlüsselung lassen sich Daten zentral poolen und verarbeiten, ohne dass diese Daten preisgegeben werden. Der Aufwand derzeitiger Verfahren ist jedoch immer noch so hoch, dass ein Einsatz für reale Probleme heute in weiter Ferne liegt.

Trusted Execution Environments

Sobald eine vertrauenswürdige zentrale Instanz existiert, werden sichere Mehrparteienberechnungen leicht umsetzbar. Es ist allerdings schwierig, ein IT-System so zu entwickeln, dass ihm alle Beteiligten vertrauen. Einen Ansatz bilden vertrauenswürdige Enklaven: Auf einem Chip unterhalb der Ebene des Betriebssystems führen sie Programme gekapselt so aus, dass niemand an die verarbeiteten Daten kommt. Die Parteien kommunizieren ver-

schlüsselt mit der Enklave. Gleichzeitig sind alle Handlungen der Enklave von den Parteien abgeschlossen. Ein aktuelles Beispiel ist die Intel SGX, die eine vertrauenswürdige Enklave realisiert, wenn man dem Chip von Intel vertraut.

Alternativ zu einer Enklave auf einem Chip ist ein Trusted Execution Environment auch in größerem Maßstab realisierbar, etwa durch einen Server in einem Rechenzentrum, der geeignet isoliert wurde. Hier sollte die Systemarchitektur, von der die Sicherheit unmittelbar abhängt, einfach nachvollziehbar sein, sodass die Sicherheit mit einem geringeren Aufwand zertifizierbar ist. Dies ebnet den Weg zu größerem Vertrauen und höherer Akzeptanz der Anwenderinnen und Anwender. Ideal wäre eine „Auditable Security", die klare und leicht überprüfbare Annahmen über die verwendeten Komponenten und den Systemaufbau formuliert. Sie setzt mehr Vertrauen in dritte Instanzen voraus als rein kryptographische Verfahren. Dennoch werden solche vertrauenswürdigen Server und das Zusammenspiel von Hardware/Software-Komponenten noch für sehr lange Zeit die wesentlich effizientere Lösung sein.

Der Algorithmus kommt zu den Daten

Eine naheliegende Alternative besteht darin, nicht die Daten zentral zu poolen, sondern das Lernende System „herumzuschicken". Es lernt zunächst auf lokal vorliegenden Daten, dann wird es an einen anderen Ort geschickt, um dort weiter zu lernen. Nach vielen Schritten wurde das Lernende System schließlich auf allen Daten trainiert. Leider ist derzeit nicht abschließend geklärt, ob dieses Verfahren die Privatsphäre ausreichend schützt. Aufgrund der in der EU-Medizinprodukteverordnung definierten Qualitätsanforderungen an Daten wird davon ausgegangen, dass das KI-System keine rekonstruierbaren personenbezogenen Daten enthält. Trotzdem könnte das Lernende System nun personenbezogene Daten beinhalten und darf deshalb nicht weitergegeben werden. So könnte es im Rahmen eines Vorher-Nachher-Vergleichs Informationen über Daten an einem bestimmten Ort preisgeben. Solange die Eigenschaften eines solchen iterativen Lernens nicht gut genug verstanden sind, kann es nicht verwendet werden. Allerdings bietet es sich an, diese Lösung mit dem Ansatz der vertrauenswürdigen Enklaven oder vertrauenswürdigen Server zu kombinieren: Das Lernende System verlässt an keinem Ort die Enklave und wird zwischen den Enklaven verschlüsselt verschickt. Dies ist wahrscheinlich der vielversprechendste Ansatz, wenn man Sicherheit und Performanz abwägt.

3.4 Sichere KI-Datenbanken

Die Verwaltung und Pflege von Datenbeständen, die für KI-basierte Analysen im Gesundheitswesen genutzt werden, müssen spezifische Anforderungen erfüllen. Diese lassen sich aus den allgemeinen Beobachtungen über Maschine-Learning-Verfahren herleiten und werden in folgendem Kapitel vorgestellt. Dazu wird zuerst die aktuelle Praxis dargestellt und auf mögliche Risiken und Probleme hin analysiert. Abschließend leiten die Autoren Handlungsoptionen ab, die die Risiken und Probleme reduzieren können.

Aktuelle Vorgehensweise

Bei KI-Systemen existieren zu jedem Analyseproblem zwei Datensätze: einer zum Trainieren und einer zum nachträglichen Überprüfen des trainierten KI-Systems. Diese Datensätze müssen repräsentativ sein. In der Trainingsphase bewerten Expertinnen und Experten die Ergebnisse. Dies geschieht gegebenenfalls in einem teilautomatisierten Verfahren. Etwaige von der Fehlerklassifikation verursachte Risiken fließen in diese Bewertung nicht ein – es wird lediglich eine Falsch-Richtig-Analyse des Outputs bei einem gegebenen Input durchgeführt. Nach abgeschlossener Trainingsphase überprüfen die Expertinnen und Experten mit Hilfe eines unabhängigen Testdatensatzes, ob für alle Daten dieses Testdatensatzes richtige Analyseergebnisse erzielt werden.

Potentielle Problemfelder und mögliche Risiken innerhalb dieses Verfahrens

- **Closed World Problem**
  Die Trainings- und Testdatensätze können notwendigerweise nur eine endliche Anzahl an Daten beinhalten. Dadurch sind möglicherweise relevante Informationen nicht enthalten („Closed World Problem", Definition siehe Infobox). Neueste Forschungsergebnisse und Beobachtungen anhand von Patientendaten können dabei helfen, die Liste der benötigten Daten fortlaufend zu aktualisieren. In der Folge ist es möglich, die verwendeten Datensätze anzupassen und gegebenenfalls zu erweitern. Im betrachteten Szenario wird dies gelöst, indem bei einem unklaren Befund oder Verdacht auf Lungenkrebs eine Computertomographie (CT) empfohlen wird. Sollte durch die CT-Aufnahmen bei einer Patientin oder einem Patienten ein Lungenkrebs identifiziert werden, so würde das KI-Assistenzsystem im nächsten Schritt eine Magnetresonanztomographie (MRT) empfehlen, um sekundäre Tumore zu entdecken. Dies wird für die meisten Patientinnen und Patienten eine richtige Empfehlung sein. Sie stellt jedoch einen systematischen Fehler für diejenigen Betroffenen dar, die großflächige Tätowierungen mit bestimmten Farbpigmenten tragen (da im MRT signifikante Verbrennungsrisiken der Haut entstehen).

Closed World Problem
Die Closed World Annahme besagt Folgendes: Alles, was nicht im Modell abgebildet ist, existiert nicht und ist somit auch nicht beobachtbar. Es wird folglich angenommen, dass außerhalb des beobachteten Datensatzes keine für die Beobachtungen relevanten Informationen existieren. In der Praxis ist diese Annahme kritisch zu betrachten – so kann de facto von keinem Verfahren ausgeschlossen werden, dass außerhalb des betrachteten Datensatzes weitere relevante Informationen existieren. Da diese aber nicht mit einbezogen werden können, unterliegt de facto jedes KI-basierte Assistenzsystem dem „Closed World Problem".

- **Potentiell verzerrte Datensätze**
  Wie in Kapitel 3.1 dargestellt, können die Trainings- und Testdatensätze (unbeabsichtigt) verzerrt oder (gezielt) manipuliert sein. Dies kann dazu führen, dass das Modell des KI-Systems auf einer fehlerhaften Datenbasis lernt und so in der Anwendung unter Umständen auch fehlerhafte Ergebnisse ausgibt.

- **Konfidenz von KI-Systemen**
  Es ist außerdem zu beachten, dass die Konfidenz von KI-Systemen (Definition siehe Infobox) nicht gleichbedeutend mit dem statistischen Begriff der Korrelation ist. Es ist möglich (und auch beobachtbar), dass verschiedene Neuronale Netze, die für die gleichen Analyseaufgaben trainiert worden sind, für widersprüchliche Analyseantworten hohe Konfidenzen angeben.

Konfidenz vs. Statistik bei KI-Systemen
Im Bereich der Künstlichen Intelligenz beziehungsweise des Maschinellen Lernens bezeichnet Konfidenz die Aufmerksamkeit des Algorithmus für bestimmte Aktionen. Aktionen, die für das Lösen der Aufgabe relevanter erscheinen, erhalten mehr Aufmerksamkeit des Algorithmus. Diese Konfidenzen entstehen während des Lernprozesses der KI-Software. Sie sind jedoch zu unterscheiden von probabilistischen Zusicherungen der Art, dass eine KI-Assistenz etwa das Erkennen eines Tumors mit einer begrenzten Fehlerwahrscheinlichkeit korrekt vorhersagt. Die Konfidenz der klassischen Statistik zeigt an, mit welcher Wahrscheinlichkeit das Ergebnis innerhalb eines bestimmten Intervalls liegt, beziehungsweise bei binären Ereignissen (Tumor – Ja oder Nein), mit welcher Wahrscheinlichkeit das Ergebnis korrekt ist. Mit dem jetzigen Stand der Technik ist es nicht möglich, die Wahrscheinlichkeit von Fehlklassifikation zu begrenzen – nicht zuletzt wegen der oben dargestellten Closed World Problematik.

Anforderungen an sichere Datenbanken

- **Nachvollziehbarkeit und Erklärbarkeit der Ergebnisse**
  Ergebnisse, die mit Hilfe eines KI-Systems ermittelt wurden, müssen nachvollziehbar und erklärbar sein. Dazu zählt vor allem, dass eine Ärztin oder ein Arzt die vorgeschlagenen Analyseergebnisse kritisch hinterfragen können muss. Wenn Medizinerinnen und Mediziner Nachfragen stellen, muss das KI-System in der Lage sein, qualifizierte und interpretierbare zutreffende Gründe für die ausgegebene Analyse darzulegen. Die behandelnden Ärztinnen und Ärzte dürfen das vorgeschlagene Ergebnis nicht unreflektiert als Ergebnis übernehmen. Vielmehr sollen sie dieses mit in die Behandlungsentscheidung einbeziehen – als einen Aspekt, der ihr Wissen und ihre Erfahrungen ergänzt.

- **Integrität der Datensätze und sichere Übertragungswege**
  KI-Systeme sind einer Klasse von potentiellen Angriffen ausgesetzt, die die Merkmale der Trainingsdatensätze verfälschen können. Es bestehen folglich hohe Anforderungen an das Sichern der Integrität der verwendeten Datensätze. Diese Anforderungen können im Fall des Gesundheitswesens aus dem IT-Sicherheitsgesetz abgeleitet werden, da die Gesundheitsversorgung eine kritische Infrastruktur im Sinne dieses Gesetzes darstellt. Kennt eine Angreiferin oder ein Angreifer das maschinelle Lernverfahren und die Testdaten, bedeutet dies ein potentielles Risiko. So könnte sie oder er möglicherweise im Rahmen von Angriffen in der ePA gezielt die entscheidenden Informationen manipulieren, also verändern oder löschen. Die KI-Software würde dann bei der Analyse mit hoher Wahrscheinlichkeit ein falsches Ergebnis ausgeben. Folglich sollten Schutzmaßnahmen nicht ausschließlich getroffen werden, um die Integrität der ePA zu wahren. Demnach müssen alle Übertragungswege zwischen der ePA und dem verwendeten KI-System sehr gut abgesichert und geschützt sein. Ergänzend sollte auch die Integrität der Datensätze vor jedem Lernschritt überprüft werden, damit etwaige Datenmanipulationen auffallen können. Bei Einführung der ePA sollte berücksichtigt werden, dass die Integrität der Datensätze eine zentrale Rolle spielt. Wie KI-Systeme vor verfälschten oder manipulierten Trainingsdaten geschützt werden können, wurde in Kapitel 3.1 aufgezeigt.

Etablierung einer staatlich beauftragten neutralen Stelle

Die Verwaltung und Pflege der KI-Analyseverfahren samt zugehörigen Trainings- und Testdatensätzen sollte an staatlich beauftragte neutrale Einrichtungen vergeben werden – so wie dies in der Europäischen Gesetzgebung für Medizinprodukte vorgesehen ist.[14] Über die oben angesprochenen notwendigen Schutzmaßnahmen hinaus sollten diese Institutionen folgende Aufgaben haben:

- **KI-Anwendungen klassifizieren:** Dieser Prozess muss die einschlägigen Normen beinhalten, die die Entwicklung und den Betrieb dieser Technik regeln. Das umfasst einerseits die Frage, ob die Anwendung als Medizinprodukt eingestuft und zugelassen wird. Andererseits sollte untersucht werden, welche Risiken mit Fehlanalysen des Systems einhergehen können.

- **KI-Anwendungen verpflichtend zertifizieren:** Die Zertifizierung muss auf den für die Risikoklassifikation relevanten und noch zu definierenden nationalen und internationalen Standards basieren (für weitere Informationen siehe Kapitel 4.1).

- **Aktualität der verwendeten Trainings- und Testdatensätze sicherstellen:** Als fehlerhaft erkannte Analyseergebnisse sowie die von der KI-Software ausgegebene Erklärung sollten der autorisierten Stelle übermittelt werden. Dafür müssen Prozesse etabliert und überwacht werden. Im nächsten Schritt sollten Expertinnen und Experten wissenschaftlich analysieren, mit welchen Maßnahmen die falschen Ergebnisse vermieden werden können. Infrage kommt, entweder zusätzliche Informationen zu verwenden

14 Und beispielsweise auch bei den Krebsregistern gehandhabt wird.

oder unzulässige statistische Korrelationen in den Trainings- und Testdatensätzen zu identifizieren und zu eliminieren. Je nach Analyseergebnis müssen neue Trainings- und Testdatensätze geschaffen werden, mit denen das KI-System nach- beziehungsweise neu trainiert oder nachgetestet wird.

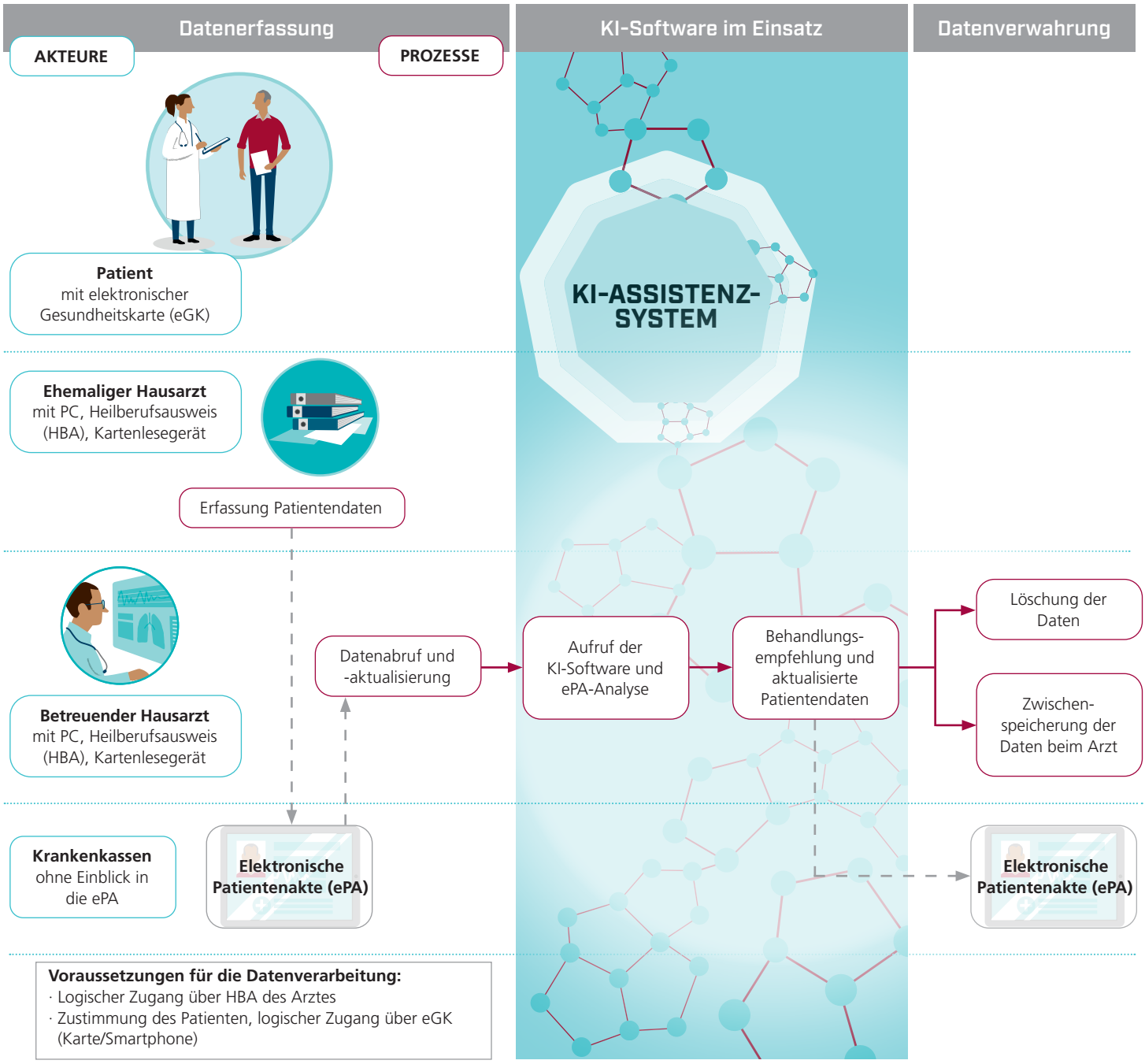
- **Rückrufprozesse für die KI-Software etablieren:** Wie auf Fehlanalysen reagiert wird, sollte von dem Schweregrad der Fehlanalyse abhängen. In schwerwiegenden Fällen sollten die staatlich beauftragten Stellen berechtigt sein, die Nutzung des KI-Systems sofort zu untersagen. Eine neue Version des KI-Systems müsste dann erneut alle Schritte der Zulassung durchlaufen („Rezertifizierung"), um für die bisherigen Nutzerinnen und Nutzer wieder freigegeben zu werden. Die staatlich beauftragten Stellen müssten außerdem analysieren, ob die bisherigen Analyseergebnisse des fehlerhaften KI-Systems aus den ePAs der Betroffenen gelöscht beziehungsweise vorläufig auf „nicht sichtbar" gestellt werden müssen, solange das Ergebnis nicht als gesichert gilt. Für das mögliche Löschen der fehlerhaften Analyseergebnisse ist ein geeignetes Verfahren zu entwickeln.

3.5 Patientendaten sicher bereitstellen

Die Einführung einer elektronischen Patientenakte – ePA (Definition siehe Infobox) – ist eine Herausforderung, die alle Sektoren des Gesundheitswesens betrifft. Gleichzeitig ist sie eine der wichtigsten Voraussetzungen dafür, dass das Anwendungsszenario „Mit KI gegen Krebs" umgesetzt werden kann. Das verdeutlicht auch die Abbildung 1 (siehe S. 29): Mit den Daten aus der ePA beginnt die Behandlung der Patientin oder des Patienten. Das folgende Kapitel beschreibt potentielle IT-Sicherheitsrisiken, die mit der Einführung der ePA einhergehen können. Zugleich stellt es mögliche Gegenmaßnahmen vor. Die skizzierten Risiken sind nicht KI-spezifisch, sondern allgemeiner Natur; die vorgeschlagenen Gegenmaßnahmen ergeben sich aus gängigen IT-Sicherheitslösungen.

Elektronische Patientenakte (ePA)
Die elektronische Patientenakte (ePA) bezeichnet die individuelle Datenbasis, in der Gesundheitsdaten der jeweiligen Patientinnen und Patienten ganzheitlich und bundesweit einheitlich gespeichert werden. Sie zielt darauf ab, Qualität, Transparenz und Wirtschaftlichkeit der Behandlungen zu verbessern. Die ePA enthält relevante medizinische Daten der Patientinnen und Patienten, von vergangenen und aktuellen Diagnosen und Befunden über Behandlungen und Operationen bis hin zu Medikationen. Dies ermöglicht eine bessere und einfachere Zusammenarbeit verschiedener Ärztinnen und Ärzte, Apothekerinnen und Apotheker, Krankenhäuser und Heilberufe. Zudem werden mögliche Risiken durch Kreuzmedikationen verringert und die Informationsverfügbarkeit in Notfällen erhöht. Der Datenschutz dieser hochsensiblen persönlichen Daten spielt eine große Rolle. Die Patientin oder der Patient soll die Hoheit über ihre/seine ePA-Daten besitzen. Die Einführung ist für alle gesetzlichen Kassen bis spätestens zum 01.01.2021 gesetzlich verpflichtend.

Abbildung 1: Sichere Datenweitergabe: Vom Patienten in die elektronische Patientenakte



Potentielle IT-Sicherheitsrisiken bei Einführung der ePA

Die flächendeckende verpflichtende Einführung der ePA ist mit zahlreichen Sicherheitsrisiken verbunden.[15] Bei der **Dateneingabe** könnten Angreiferinnen und Angreifer die Daten beim Upload abfangen und gegebenenfalls in veränderter Form weiterleiten. Darüber hinaus wäre es möglich, dass Unbefugte falsche Patientendaten in die ePA einspeisen. Um diese Risiken auszuschließen, bedarf es eines kombinierten Sicherheitsansatzes: So sollte erstens der Datenverkehr verschlüsselt, integritäts- und authentizitätsgesichert sein. Zweitens sollte stets eine Authentisierung der Kommunikationspartner stattfinden, sodass nur autorisierte Personen Zugriff auf das System haben und Manipulationen mittels kryptographischer Verfahren wie zum Beispiel Verschlüsselung und Signaturen verhindert werden. Abbildung 1 zeigt die sichere Datenweitergabe – von der Patientin oder des Patienten in die elektronische Patientenakte (siehe S. 29). Die zentralen Akteure sind die Patientinnen und Patienten, die oder der ehemalige oder aktuelle Hausärztin und Hausarzt sowie die Gesundheitskasse. Die Ärztin oder der Arzt erhält über eine 2-Faktor-Authentisierung Zugang zu den ePA-Daten. Sie oder er benötigt hierzu die elektronische Gesundheitskarte (eGK) der Patientin oder des Patienten, ihren oder seinen Heilberufsausweis (HBA) sowie die zugehörige PIN. Dadurch ist zwar der Datenzugang abgesichert, aber es sind noch nicht alle Risiken beseitigt. Wenn nach dem Upload in die ePA sensible Daten auf dem Gerät der Ärztin oder des Arztes verbleiben und dort unzureichend geschützt sind (z. B. in Form einer unverschlüsselten Festplatte), könnten sie möglicherweise gestohlen werden. Die Lösung: Das sofortige, zuverlässige Löschen nicht mehr benötigter Daten beseitigt dieses Risiko; gleichzeitig mindert die Verwendung eines verschlüsselten Dateisystems den Druck, diese nicht mehr benötigten Dateien (zeitnah) zuverlässig zu löschen. Auch wäre unter dem Vorbehalt vorhandener regulatorischer Rahmenbedingungen zu erwägen, ob diesem Risiko in zentral zur Verfügung gestellten Multi-Cloud-Lösungen nicht besser begegnet werden kann.

Sind die Daten erfolgreich in die ePA übertragen worden, kann es auch bei der **Datenvorhaltung** (also dem Speichern von Daten) zu IT-sicherheitsrelevanten Ereignissen kommen. Ein Beispiel ist das unbefugte Abrufen und auch Kopieren von patientenbezogenen Daten aus der ePA. Um diesem entgegenzuwirken, müssen die Datensätze in verschlüsselter Form aufbewahrt werden; die Entschlüsselung darf nur mit der eGK oder dem HBA in Kombination mit einer Authentisierung der Kommunikationspartner möglich sein. Darüber hinaus können Verfahren zur Anomaliedetektion beim Abrufverhalten eingesetzt werden – sie würden beispielsweise warnen, wenn eine Orthopädiepraxis Leberwerte abruft. Verwahren die Patientinnen und Patienten oder Ärztinnen und Ärzte die eGK und den HBA gemeinsam mit der PIN auf, können diese Informationen gestohlen werden. Fortlaufend aktualisierte Sperrlisten für die eGKs sowie die HBAs können dieses Risiko erheblich reduzieren. Zusätzlich könnten biometrische Authentisierungsverfahren angewendet werden. Diese Vorschläge greifen allerdings nur bedingt, wenn eine Ärztin oder ein Arzt ge-

15 Ziel des Papiers ist es, einen Rahmen für die Umsetzung des Anwendungsszenarios „Mit KI gegen Krebs" im Jahr 2024 zu entwickeln. Eine Analyse der Frage der aktuellen rechtlichen und technisch-organisatorischen Vorgaben zur Einführung der ePA kann an dieser Stelle nicht geleistet werden, kann jedoch in Folgepublikationen Eingang finden.

zwungen wird, die HBA herauszugeben, etwa durch Erpressung oder unter Androhung physischer Gewalt. In diesem Fall ist kein umfassender Schutz möglich – unterstützend könnte jedoch eine Sperre wirken, die eintritt, sobald anomale Abrufaktivitäten identifiziert werden. Diese könnte zum Beispiel bei einem Abruf von sehr vielen Patientendaten einen Hinweis geben.

Auch bei der **Datenausgabe** an die behandelnde Fachärztin oder den behandelnden Facharzt könnten Unbefugte Daten abrufen. Hinzu kommt die Möglichkeit, dass die Daten während des befugten Abrufs aus der ePA kopiert und (gegebenenfalls verfälscht) weitergeleitet werden könnten. Hier bieten sich die gleichen Lösungen an wie im Fall der Dateneingabe: ein verschlüsselter, integritäts- und authentizitätsgesicherter Datenverkehr in Verbindung mit einer Authentisierung der Kommunikationspartner.

Im Anwendungsszenario „Mit KI gegen Krebs" können Patientinnen und Patienten im Nachgang der Behandlung ihre Daten freiwillig und geschützt weitergeben und so zu einer kontinuierlichen Verbesserung von KI-basierten Diagnosen und Therapien beitragen. Bei der **freiwilligen und geschützten Datenfreigabe** könnte eine ungenügende Anonymisierung der Daten dazu führen, dass einzelne Daten rekonstruiert und erneut zusammengeführt werden könnten. Dies ließe wiederum Rückschlüsse auf die Identität der Patientin oder des Patienten zu. Mit der Wahl angemessener Anonymisierungs- oder Pseudonymisierungsverfahren (Definition siehe Infobox Seite 14) lässt sich dieses Risiko minimieren. In Ergänzung dazu soll die IT-Infrastruktur/Software so gestaltet werden, dass bei Bedarf weitere Anonymisierungsschritte möglich sind.

Verfügbarkeit vs. Privacy

Patientendaten, wie etwa Angaben zu einer Medikamentenunverträglichkeit, sind personenbezogene Daten und müssen gut geschützt werden. Im Zweifelsfall ist für eine Patientin oder einen Patienten aber noch wichtiger, dass diese Daten verfügbar sind, wenn sie benötigt werden. Insbesondere in einem Notfall sollten Schutzmechanismen den Zugang zu relevanten Daten nicht behindern, sondern im Sinne eines berechtigten Interesses nutzbar sein.[16] IT-Sicherheit für die Medizin sollte die Verfügbarkeit von Daten klar priorisieren. Die Sicherheitsmechanismen sollten an erster Stelle darauf abzielen, unerlaubte Zugriffe zu entdecken und Täterinnen und Täter zu identifizieren. Die „Break-Glass-Metapher" verdeutlicht die angestrebte Richtung: Im Notfall kann das Glas zerbrochen werden, der Zugriff ist dann uneingeschränkt möglich.

16 Die DSGVO sieht als einen Erlaubnistatbestand der Datenverarbeitung das sogenannte berechtigte Interesse der verarbeitenden Stelle an der jeweiligen Datenverarbeitung vor. In dem Fall ist das berechtigte Interesse gegeben, da z. B. die Ärztin oder der Arzt und Rettungssanitäterinnen oder -sanitäter im Sinne des Wohls der Patientinnen und Patienten und zur Erfüllung ihrer Dienstpflicht (heilen, Leben retten etc.) auf die entsprechenden Daten zugreifen können sollten.

3.6 KI-Systeme sicher in den klinischen Prozess integrieren

Bei der Integration eines KI-gestützten Assistenzsystems in den klinischen Prozess können diverse, insbesondere IT-bedingte Risiken entstehen, die berücksichtigt und gesondert adressiert werden müssen. Einige potentielle Probleme und mögliche Lösungen diskutiert der folgende Abschnitt anhand eines konkreten Beispiels.

Beispiel: Berechnungs-Assistent für die Wirkstoffkombination der Chemotherapie

Training auf unvollständigen Datensätzen: Das Machine-Learning-Modell (ML-Modell), das relevante Erkenntnisse wie etwa Leitlinien und Studien analysieren soll, wurde nur mit einem kleinen Teil relevanter klinischer Daten trainiert, zum Beispiel mit Krankheitsklassifikationen durch Patientenmetadaten wie ICD-Codes.[17] Wichtige Details, etwa aus molekulargenetischen Laborparametern hat das System jedoch nicht berücksichtigt, weil der Zugang zu den entsprechenden Daten fehlte. Die Stabilität des ML-Modells basiert folglich auf stark abstrahierten Klassifizierungen. Das kann zu wenig hilfreichen Prognosen führen. Eine mögliche Lösung für dieses Problem: den Zugang zu Studiendaten verbessern, etwa mittels Open-Data- und Open-Access-Initiativen. Darüber hinaus würde eine bessere Nachvollziehbarkeit und Erklärbarkeit von Ergebnissen erreicht werden, die die Aussagekraft des KI-gestützten Assistenzsystems zusätzlich stärken. Diese Fragen sind aktuelle und relevante Forschungsfelder.

Veraltete Versionierung und Aktualität: Im besten Fall behebt jede neue Version des ML-Modells Fehler in der Prognose und sorgt so für präzisere Vorhersagen. Wie kann sichergestellt werden, dass alle Anwenderinnen und Anwender dieses Modells jeweils die aktuelle Version nutzen? Hier könnte ein zentrales Verzeichnis mit versionierten ML-Modellen helfen, wie es in der Europäischen Gesetzgebung verbindlich vorgeschrieben ist. Dieses könnte die Grundlage für automatisierte Update-Mechanismen sein, die das jeweils aktuelle Modell zugänglich machen. Gleichzeitig muss aber Anwenderinnen und Anwendern des Modells auch klar angezeigt werden, dass eine neue Version verwendet wird, zum Beispiel mit der Versionsnummerierung und dem Updatedatum. So ist es ihnen möglich, Ergebnisse von gestern und heute zu vergleichen oder jüngste Prognosen gegebenenfalls zu wiederholen. Die Nutzerin oder der Nutzer des ML-Modells müssen auch die Möglichkeit haben, im Zweifelsfall auf mehrere Versionen zugreifen und deren Ergebnisse vergleichen zu können. Es sollte sichergestellt werden, dass durch ein automated deployment der jeweils aktuellen Version die bisher identifizierten Fehlklassifizierungen eliminiert werden.

17 ICD bedeutet „International statistical Classification of Diseases and related health problems". ICD-Codes sind von der WHO festgelegte Codes, welche der internationalen Klassifikation sowie Einordnung von Krankheiten und Medikamenten dienen.

Sicherung vor falschen Updates: Wenn das ML-Modell aktualisiert wird, sobald es sich verbessert hat, ist dies ein potentielles Einfallstor für Angreiferinnen und Angreifer. Das ML-Modell könnte nämlich bewusst durch ein Modell ersetzt werden, das für eine spezielle Patientin oder für einen speziellen Patienten manipulierte Ergebnisse liefert. Es stellt sich deshalb die Frage, wie die Integrität der ML-Modelle auf dem Weg zu Nutzerinnen und Nutzern im klinischen Umfeld sichergestellt werden kann. Eine Lösung könnte ein sogenanntes Trusted Execution Environment sein. Hierbei ist der Zugang zu den Modellen nur autorisierten Nutzerinnen und Nutzern möglich. Darüber hinaus sollten unabhängige, automatisierte Software-Testverfahren sicherstellen, dass neue Versionen des Modells keine schlechteren Ergebnisse liefern. Hierzu ist ein geeigneter Freigabeprozess mit klaren Qualitätskriterien[18] zu etablieren.

Problematische Priorisierung von Ergebnissen: Das KI-gestützte Assistenzsystem muss entscheiden, in welcher Reihenfolge verschiedene Therapieoptionen dargestellt werden. Die Autoren dieses Papiers gehen davon aus, dass Ärztinnen und Ärzte sowie Patientinnen und Patienten im klinischen Alltag nicht alle Ergebnisse, sondern nur die vermeintlich relevantesten anschauen werden. Zentral ist deswegen eine geeignete Sortierung der Ergebnisse. Wie könnte sie aussehen?

Derzeit werden Systeme erforscht, die Wissen dynamisch auswerten und priorisieren. Weil es sich dabei aber um unüberwachtes Lernen handelt, existiert derzeit keine rechtliche Möglichkeit, diese als Medizinprodukte zuzulassen. Hier könnten ähnliche selbstverstärkende Effekte entstehen wie bei Suchmaschinenergebnissen: Ein KI-Algorithmus bestimmt die Reihenfolge. Er lernt aus dem Verhalten der Suchenden, die häufig zuerst schwerwiegende Krankheitsbilder auswählen. In der Folge schlägt der Algorithmus auf der ersten Seite diese schweren Krankheiten vor, obgleich die Symptome auch bei harmlosen Erkrankungen auftreten. Darüber hinaus werden bestimmte Ergebnisse aufgrund der persönlichen Präferenz gar nicht mehr angezeigt. Das eigene Verhalten kann also die Unabhängigkeit der Ergebnisse stark einschränken.

In ähnlicher Weise könnte ein KI-gestütztes Assistenzsystem im Gesundheitswesen vorrangig schwerwiegende Erkrankungen detektieren und eher harmlose weniger berücksichtigen. Die Folge: kostspielige Untersuchungen und ein Anstieg von schwerwiegenden Diagnosen, selbst wenn sich diese später nicht bestätigen. Aus diesen Gründen sollten selbstverstärkende Effekte bei der Ergebnisdarstellung untersucht und korrigiert werden. Wichtig ist es, der Software-Ergonomie Rechnung zu tragen. Sie stellt folgende Fragen in den Mittelpunkt: Wie nutzen Anwenderinnen und Anwender die Software? Wie kritisch gehen sie mit den Ergebnissen um? In Usability-Tests sollte überprüft werden, ob Anwenderinnen und Anwender die Ergebnisse und die Ergebnisdarstellung reflektiert wahrnehmen. Rückmeldungen von Nutzerinnen und Nutzern während der Ergebnisdarstellung

18 Bei der Entwicklung von Qualitätskriterien müssen unter Umständen konfligierende Ziele und Risiken sowie Vor- und Nachteile gegeneinander abgewogen werden. So könnte zum Beispiel unter bestimmten Bedingungen eine Verbesserung des medizinischen Nutzens mit der Qualität des Datenschutzes im Konflikt stehen. Die Qualitätskriterien, die zum Einsatz kommen sollen, müssen daher unter Einbindung aller relevanten Anspruchsgruppen im Rahmen eines gesellschaftlichen Diskurses entwickelt werden (siehe hierzu auch 4.2).

aktiv einzuholen, ist daher relevant (beispielsweise „dieses Ergebnis passt für meine Anfrage nicht, weil …“). So können Entwicklerinnen und Entwickler das Feedback berücksichtigen: Das KI-System lernt also von den Nutzerinnen und Nutzern. Ebenso sollen auch die Anwenderinnen und Anwender von den Erfahrungen des KI-Systems lernen. Die Darstellung der Ergebnisse sollte sowohl quantitative Resultate (etwa „85 Prozent der Nutzerinnen und Nutzer fanden dies relevant“) als auch qualitative Hinweise enthalten („nach dieser Prognose haben Nutzerinnen und Nutzer folgende Abfragen getätigt …“). So entsteht ein übergreifendes Gemeinschaftswissen, auf das jede und jeder Einzelne Zugriff hat. Alle können davon profitieren, würden aber gleichzeitig auch selbst dazu beitragen.

4. Gestaltungsoptionen für sichere KI-Systeme in der Medizin

Dieses Whitepaper hat bislang die Herausforderungen beim Einsatz von KI-Anwendungen im Gesundheitswesen aufgezeigt sowie technische Lösungen für diese vorgeschlagen. Die technischen Lösungen können jedoch nur so gut funktionieren, wie ein Rechtsrahmen dies zulässt. Das folgende Kapitel stellt deshalb zunächst regulatorische Gestaltungserfordernisse und -optionen dar. Anschließend thematisiert es offene Fragen, die in einem intensiven gesellschaftlichen Diskurs erörtert werden sollten.

4.1 Regulatorische Gestaltungserfordernisse und -optionen

- **Gemeinsame Leitlinien und Prüfvorschriften für die Zulassung und Zertifizierung entwickeln:** Dynamische Softwarearchitekturen führen dazu, dass die Funktion und Wirkungsweise eines Medizinproduktes vor Inverkehrbringung heute nicht mehr mess-, beleg- und im Bedarfsfall zertifizierbar ist. Dies trifft auch auf (weiter-)Lernende Systeme zu. Daraus abzuleiten, dass ein Produkt bei jedem Software-Update als neues Produkt zu betrachten wäre, erscheint nicht zielführend. Gleichwohl sollte der Zulassungsprozess weiterentwickelt werden. Neben dem Produkt selbst ist es auch notwendig, dessen Betrieb und Zertifizierungsanforderungen für Updates zu betrachten. Aus diesem Spannungsfeld ergeben sich konkrete Maßnahmen: Der Gesetzgeber sollte gemeinsam mit den betroffenen Stakeholdern Leitlinien sowie Prüfvorschriften für und Anforderungen an einen Zulassungsprozess und damit verbunden eine Zertifizierung der KI-Systeme[19] erarbeiten. Bereits existierende Richtlinien, wie beispielsweise die DIN EN ISO 13485[20], die ISO 14971[21] sowie IEC 62304, sollten hierbei berücksichtigt werden. So kann sichergestellt werden, dass nur sichere und robuste KI-Systeme eingesetzt werden. Zudem sollte eine Kennzeichnungspflicht der eingesetzten KI-Algorithmen eingeführt werden, die deren Eigenschaften und Zulassung für bestimmte Anwendungsgebiete transparent macht.

- **Gemeinsame Leitlinien und Prüfvorschriften für die Zulassung und Zertifizierung der Betreiber der KI-Datenbanken entwickeln:** Der Gesetzgeber sollte gemeinsam mit den relevanten Stakeholdern ebenfalls Leitlinien sowie Prüfvorschriften und Anforderungen an einen Zulassungs- und Zertifizierungsprozess für die zertifizierten Betreiber der KI-Datenbanken entwickeln.

19 Hierbei wäre u.a. zu unterscheiden, ob es sich bei dem System um ein ‚ausgelerntes' oder ‚weiterlernendes' System handelt. Allerdings ergeben sich in beiden Fällen Anforderungen an den Betrieb und den Betreiber des KI-Systems, die im Abschnitt 3.4 diskutiert werden.

20 In dieser Norm sind Anforderungen zur Herstellung und Einführung von Medizinprodukten festgeschrieben. Es handelt sich um Anforderungen für ein umfassendes Managementsystem in Bezug auf die Herstellung und das Design medizinischer Produkte.

21 Risikomanagement medizinischer Produkte: Das Risikomanagement medizinischer Produkte muss als eine zwingende Voraussetzung für ein Medizinprodukt angesehen werden. Diese Norm umfasst neben der Risikoanalyse auch die Risikobewertung sowie die Risikobeherrschung durch ein Maßnahmen-Management bis hin zu einer Neubewertung.

- **Hersteller gesetzlich zur Mängelbehebung verpflichten:** Es entstehen neue, und wenn nötig, strengere Sicherheitsanforderungen an die Anwendungen, die im Rahmen der Zulassung erfüllt werden müssen. Für den Betrieb eines KI-Systems ist dies in der europäischen Gesetzgebung geregelt und fest definiert. Bestimmte Produkteigenschaften können vor Markteinführung überprüft und bewertet werden. Darüber hinaus sollten nachgelagert auch Funktionsstörungen beobachtet und von den Herstellern im Sinne der Mängelbehebung behoben werden – unabhängig davon, ob sie nur auf KI-Funktionalitäten oder sonstige Systemanpassungen zurückzuführen sind. Hierfür sollte geprüft werden, ob im Medizinproduktegesetz beziehungsweise den zugrundeliegenden EU-Verordnungen (inkl. deren zukünftiger Anpassungen) eine **„Duty of Care"-Verpflichtung der Hersteller** entsprechender IT-/KI-Systeme oder -komponenten vorgesehen ist. Ist dies nicht der Fall, sollte diese Verpflichtung etabliert werden. Diese Weiterentwicklung des bestehenden **Vigilanzsystems** (Beobachtungs- und Meldesystem) sollte die Daten- und IT-Sicherheit sowie die Manipulationssicherheit von Systemen regeln und garantieren. Hierfür sollten auch die erforderlichen Gesetzesnovellen intensiver an den spezifischen Eigenschaften von IT-Systemen ausgerichtet sein. Hersteller wären also viel stärker als bisher dazu angehalten, nicht nur Produkte zurückzurufen, sondern auftretende Mängel (z. B. in Form von Sicherheitslücken) durch entsprechende Updates zu beheben.

- **Unabhängige autorisierte Betreiber des KI-Assistenzsystems einsetzen:** Diese staatlich beauftragten neutralen Einrichtungen sollten damit beauftragt werden, die Analyseverfahren und Datensätze zu verwalten und zu pflegen. Diese Einrichtungen dürfen nicht dazu befugt sein, Daten zu verändern oder einzuspeisen, da ein eigenes ökonomisches Interesse vorliegen könnte.

- **Ein unabhängiges Prüfkomitee einsetzen:** Ein interdisziplinäres Expertengremium sollte in regelmäßigen Abständen die Funktionsweise der zertifizierten und eingesetzten KI-Systeme überprüfen. Es wäre sinnvoll, dieses Komitee beim Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) einzurichten. Ferner sollten bei den Herstellern **Rückrufprozesse** etabliert werden, um im Falle des Nicht-Funktionierens eines Systems handlungsfähig zu sein.

Daten sind die Grundlage der Digitalisierung, ohne sie sind KI-Systeme nicht funktionsfähig. Die Gesellschaft wird von den Potentialen KI-basierter Systeme im Gesundheitswesen nur dann profitieren, wenn genügend nutzbare Daten verfügbar sind. Gleichzeitig sind vor allem medizinische Daten sehr sensible Daten, die eines besonderen Schutzes bedürfen. Deshalb ist für den Einsatz von KI-Systemen im Gesundheitswesen ein risikobezogener Ansatz zielführend und angemessen. Die oder der Einzelne sollte über die Nutzung der eigenen Gesundheitsdaten souverän bestimmen können.[22] Die eGK, der HBA und die

22 Hierbei ist zu beachten, dass es neben der Datenerhebung auf Einwilligungsbasis auch gesetzliche Ermächtigungstatbestände gibt (siehe z. B. DSGVO), die Ausnahmen darstellen.

ePA sind zentrale technische Instrumente dafür. Vor dem Hintergrund der DSGVO sowie der Bestimmungen für Berufsgeheimnisträger ergeben sich daher folgende Gestaltungsoptionen:

- **Krankenkassen sollten Sperrlisten führen:** Als ausgebende Stellen der eGKs und der HBAs sollten die Krankenkassen **Sperrlisten** führen, um einen unautorisierten Zugriff auf Daten zu verhindern. Diese Listen müssen fortlaufend aktualisiert werden, sodass im Fall des Verlusts die jeweilige Berechtigungskarte wertlos ist. Der Sperr-Notruf 116 116 könnte um die eGK und die HBA erweitert werden. Zu erwägen ist auch eine entsprechende Selbstverpflichtung der Krankenkassen, sich an dem System zu beteiligen.

- **Rückfall-Lösung einführen:** Eine Rückfall-Lösung könnte die Sperrung der eGK ergänzen. Bei ihr handelt es sich um einen Modus, in dem der Funktionsumfang eingeschränkt ist, aber die wichtigsten Funktionen eines Systems aufrechterhalten bleiben können. Dies gilt auch neben dem Verlust der eGK für viele andere Fälle, wie beispielsweise:
  - Die Patientin oder der Patient ist gesundheitlich nicht in der Lage, die eGK und die PIN zu benutzen.
  - Eine Störung der Telematik-Infrastruktur liegt vor; diese kann die komplette Infrastruktur oder Teile betreffen.
  - Die ePA-Daten können nicht abgerufen werden.
  - Der Server des KI-Software-Betreibers ist nicht betriebsbereit.
  - Die Patientin oder der Patient hat die PIN vergessen, da diese zu selten angewendet wird.
  - Die Ärztin oder der Arzt hat den HBA durch einen Kleidungswechsel zu Hause vergessen.

- **Mindestanforderungen an die Sicherheit der Dateninfrastrukturen und der Rechenzentren formulieren:** Die Autoren gehen davon aus, dass keine weitere spezifische IT-Sicherheitsregulierung (NIS/IT-SIG oder Ähnliches) benötigt wird. Die IT-Infrastrukturen, die für eine Umsetzung des Anwendungsszenarios gebraucht werden, unterliegen aktuell bereits dem Anwendungsbereich geltender Rechtssetzung. Dies schließt allerdings nicht aus, dass der Gesetzgeber die einschlägigen Rechtsverordnungen anpassen kann, indem er Mindestanforderungen definiert. Diese Mindestanforderungen sollten festlegen, dass die Daten nur innerhalb der Europäischen Union gespeichert und verarbeitet werden dürfen. Dabei ist in einem ersten Schritt wichtig, dass die im Szenario beschriebenen KI-Systeme und die damit verbundenen Infrastrukturen auch von der BSI-Verordnung für kritische Infrastrukturen (KRITIS-VO) systematisch erfasst werden. In einem zweiten Schritt sind auch entsprechende Sicherheitsanforderungen für den Aufbau der notwendigen KI-Systeme festzulegen.

- **Eine forschungskompatible elektronische Patientenakte (ePA) einführen:** Damit Patientinnen und Patienten ihre Datensätze nach der Behandlung der (universitären) Forschung zur Verfügung stellen und KI-Methoden weiterentwickelt werden können, bedarf es einer forschungskompatiblen ePA. Das bedeutet, dass die relevanten Daten in einer hohen Qualität, vollständig und in einer weiterverwendbaren Form vorliegen sollten.

- **Elektronische Patientenakte (ePA) zu einer erweiterten elektronischen Patientenakte (eePA) erweitern:** Vor allem bei der Vorsorge werden weitere Patientendaten benötigt, um Patientinnen und Patienten statistisch verlässlich zu möglichen Risikogruppen zuordnen zu können. Benötigt werden Informationen wie Anamnese, Allergien, Wellness-Maßnahmen, Diät-Programme, Krankengymnastik und andere Daten. Mit dieser Erweiterung hin zu einer erweiterten elektronischen Patientenakte (eePA) würden deutlich mehr Akteure als bislang einbezogen werden. Diese Akteure sollten mindestens einen zeitlich und inhaltlich begrenzten sicheren Zugriff auf diese Daten erhalten, sofern Patientinnen und Patienten entsprechend der gesetzlichen Vorgaben zustimmen.

- **IT-Sicherheitsprobleme weiter erforschen:** Nicht alle beschriebenen IT-Sicherheitsprobleme, die beim Einsatz von KI-Systemen im Gesundheitswesen auftreten könnten, können mit den aktuell zur Verfügung stehenden technischen Lösungen beantwortet werden. Deshalb ist die Wissenschaft aufgerufen, **diese Probleme zu erforschen** und möglichst zuverlässige Lösungen zu entwickeln. Die politischen Entscheidungsträgerinnen und Entscheidungsträger in Bund und Ländern sollten passende Programme ins Leben rufen und die entsprechende Forschungsförderung bereitstellen.

4.2 Gesellschaftsrelevante Fragen

Zentrale Sicherheitsprobleme von KI-Systemen werden sich zukünftig häufig technisch lösen lassen, etwa mit Hilfe einer Vollständigkeitsprüfung von Daten oder – bis zu einem bestimmten Grad – mit Hilfe der Zertifizierung der technischen Komponenten. Die offenen gesellschaftsrelevanten Fragen müssen aber in einem breiten, ergebnisoffenen gesellschaftlichen Diskurs[23] erörtert und beantwortet werden. Einbezogen werden sollten alle betroffenen Stakeholder: vor allem Patientinnen und Patienten, Angehörige, Ärztinnen und Ärzte, Ethikerinnen und Ethiker, Rechtswissenschaftlerinnen und -wissenschaftler, politische Entscheidungsträgerinnen und Entscheidungsträger. Fragen, die in einer möglichst breiten politischen und gesellschaftlichen Diskussion erörtert werden sollten, skizziert der folgende Abschnitt.

23 Siehe auch Deutscher Ethikrat 2017.

Dateninfrastruktur betreiben, warten und pflegen: Im vorliegenden Whitepaper wird davon ausgegangen, dass Patientinnen und Patienten mithilfe der eGK souverän über ihre Daten entscheiden können. Die daran angeschlossene ePA bildet eine Schnittstelle zwischen Patientinnen und Patienten, behandelnden Ärztinnen und Ärzten und den KI-Systemen. Das Anwendungsszenario „Mit KI gegen Krebs" konkretisiert allerdings nicht, wo und in welcher Form Daten (zwischen-)gespeichert, übertragen und erweitert werden. Dies betrifft sowohl die elektronischen Patientendaten als auch die Meta-Daten der ePA, die die KI-Software bewertet hat. Verteilte Cloud-Infrastrukturen könnten einen Lösungsansatz darstellen, da diese im Rahmen bestehender KRITIS-Regulierungen bereits größtenteils abgedeckt sind. Offen ist, wer die dafür notwendige Infrastruktur bereitstellt, sie wartet und pflegt. Im gesellschaftlichen Diskurs sollten die folgenden Fragen beantwortet werden:

Wen möchten wir als Gesellschaft mit der Bereitstellung, Wartung und Pflege verteilter Cloud-Infrastrukturen betrauen? Welche Sicherheitsstandards müssen verteilte Cloud-Infrastrukturen erfüllen?

KI-Assistenzsystem bereitstellen und betreuen: Zu klären sind außerdem Fragen, welche die operative Umsetzung betreffen. Zu klären ist, welche Institutionen das KI-System finanzieren, pflegen, kontinuierlich trainieren und auf Anfrage einer Ärztin oder eines Arztes die neueste KI-Software zur Verfügung stellen können. Wie unter 4.1 gefordert, müssen diese Institutionen unabhängig sein und dürfen über keine eigenen Einspeise- oder Veränderungsmöglichkeiten verfügen. Folgende Fragen sollten deshalb beantwortet werden:

Wer finanziert und pflegt die KI-Software? Wer ist dafür zuständig, aktuelle Versionen bereitzustellen? Welchen Institutionen bringen wir so viel Vertrauen entgegen?

Nutzen und Risiko abwägen: Ebenso wie viele andere medizinische Methoden der Diagnostik und Therapie werden auch KI-Assistenzsysteme nicht ohne Risiken eingesetzt werden können. In der Diagnostik können falsch-positive und falsch-negative Ergebnisse zu Fehlbehandlungen und schweren physischen, psychischen und finanziellen Belastungen führen. Derartige Risiken werden sich nicht vollständig ausschließen lassen. Genau aus diesem Grund werden Arzneimittel und Medizinprodukte bereits streng reguliert. Mit Künstlicher Intelligenz könnten neue Nutzen-Risiko-Abwägungen notwendig werden. So könnten mit Hilfe von Big-Data-Analysen insgesamt mehr Krankheiten früher entdeckt werden. Dies könnte aber auch mit dem Risiko von mehr falsch-positiven Befunden einhergehen. Deshalb sollte in einem gesellschaftlichen Diskurs die Frage geklärt werden:

Unter welchen Umständen und bis zu welcher Höhe sind wir bereit, als Gesellschaft „Fehlerquoten" zu akzeptieren, wenn auf der anderen Seite hoher medizinischer Nutzen geschaffen werden kann?

Verwenden von Daten (Zweckgebundenheit): Wie Patientinnen und Patienten den Zugriff auf ihre Daten und deren weitere anonymisierte beziehungsweise pseudonymisierte Verwendung (zum Beispiel für Forschungsprojekte) autorisieren sollten und können, bedarf weiterer Ausgestaltung und Konkretisierung. Gesellschaftlich erörtert werden sollten daher unter anderem folgende Fragen:

Welche Daten sollen Patientinnen und Patienten weitergeben (können)? Wie eng sollte die Zweckgebundenheit einer freiwilligen und geschützten Datenfreigabe bei explorativer Forschung ausgelegt werden?

Verantwortung und Haftung: Auch die Regelung von Verantwortung und Haftung stellt eine zentrale Frage dar, die gesellschaftlich adressiert werden sollte. Grundsätzlich sollte der Mensch Letztentscheider bleiben – so wie im Anwendungsszenario der Patient Herr Merk, den seine Ärztinnen und Ärzten zuvor umfassend informiert haben. Aber auch dann könnten falsch verarbeitete Informationen möglicherweise schwerwiegende Behandlungsfehler verursachen, etwa bei einer Operation. Zu diskutieren ist, wer die Verantwortung für Fehler trägt und ob der Einsatz von KI-Systemen im Sinne einer Haftung versicherbar sein sollte. Die zentralen Fragen lauten deshalb:

Wie sollte die Verantwortung und Haftung zwischen dem Betreiber des KI-Systems und dem medizinischen Personal aufgeteilt werden? Wie kann erreicht werden, dass die Patientin oder der Patient eine bestmöglich informierte Letztentscheidung trifft, die die medizinischen Chancen und Risiken miteinbezieht, ohne dass die Verantwortung einseitig auf ihn abgewälzt wird?

Transparenz der Ergebnisse, Nachvollziehbarkeit versus Erklärbarkeit: Je komplexer ein KI-Verfahren, desto intransparenter werden die Berechnungsschritte, mit Hilfe derer das KI-System zum Ergebnis kommt. Für Anwenderinnen und Anwender heißt das: Sie können die Rechenwege der KI schwerer nachvollziehen. Dies birgt die Gefahr, dass sie korrekte Ergebnisse falsch interpretieren. Ebenso denkbar ist, dass Anwenderinnen und Anwender verzerrte oder manipulierte Ergebnisse unbemerkt verwenden. Für eine korrekte Behandlung ist es nicht nur wichtig zu wissen, ob eine bestimmte Krankheit richtig diagnostiziert wurde. Die Ärztin oder der Arzt sollte auch erfahren, warum die Krankheit diagnostiziert wurde. Die Schwierigkeit: So wünschenswert eine maximale Nachvollziehbarkeit einerseits erscheint, könnte sie andererseits zu einer Informationsüberflutung führen. Weder den Patientinnen und Patienten noch dem medizinischen Personal wäre dadurch geholfen. Dieses Spannungsfeld gilt es im gesellschaftlichen Diskurs auszutarieren. Die Fragen lauten deshalb:

Wie viel Auskunftsrecht über die Berechnung eines KI-Systems müssen Ärztinnen und Ärzte sowie Patientinnen und Patienten haben? Welche Regeln sollte der Gesetzgeber für die Nachvollziehbarkeit und Erklärbarkeit von KI-basierten Medizinprodukten schaffen?

Über dieses Whitepaper

Autoren
Die folgenden Autoren sind Mitglieder der Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme:

Prof. Dr. Jörn Müller-Quade, Karlsruher Institut für Technologie
Prof. Dr. Werner Damm, Universität Oldenburg
Prof. Dr. Thorsten Holz, Ruhr-Universität Bochum
Dr. Detlef Houdeau, Infineon Technologies AG
Thomas Schauf, Deutsche Telekom AG
Prof. Dr. Werner Schindler, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Die nachfolgenden Autoren sind Mitglieder der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege der Plattform Lernende Systeme.

Prof. Dr. Thomas Neumuth, Universität Leipzig
Dr. Matthieu Schapranow, Hasso-Plattner-Institut

Redaktion
Stephanie Dachsberger, Geschäftsstelle der Plattform Lernende Systeme
Dr. Thomas Schmidt, Geschäftsstelle der Plattform Lernende Systeme
Eva Bräth, Geschäftsstelle der Plattform Lernende Systeme
Dr. Ursula Ohliger, Geschäftsstelle der Plattform Lernende Systeme

Über die Plattform Lernende Systeme

Lernende Systeme im Sinne der Gesellschaft zu gestalten – mit diesem Anspruch wurde die Plattform Lernende Systeme im Jahr 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Fachforums Autonome Systeme des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften initiiert. Die Plattform bündelt die vorhandene Expertise im Bereich Künstliche Intelligenz und unterstützt den weiteren Weg Deutschlands zu einem international führenden Technologieanbieter. Die rund 200 Mitglieder der Plattform sind in Arbeitsgruppen und einem Lenkungskreis organisiert. Sie zeigen den persönlichen, gesellschaftlichen und wirtschaftlichen Nutzen von Lernenden Systemen auf und benennen Herausforderungen und Gestaltungsoptionen.

Literatur

Angwin et al. (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (abgerufen am 29.08.2019).

Ärzteblatt GmbH (2019): Künstliche Intelligenz diagnostiziert genauer als (unerfahrene) Kinderärzte. https://www.aerzteblatt.de/nachrichten/101056/Kuenstliche-Intelligenz-diagnostiziert-genauer-als-(unerfahrene)-Kinderaerzte (abgerufen am 29.08.2019).

Beck et al. (2019): Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf (abgerufen am 29.08.2019).

BSI/ANSSI (2019): Deutsch-französisches IT-Sicherheitslagebild, 2. Edition. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/DE-FR-Lagebild/de-fr_Lagebild.pdf;jsessionid=76F139A8135A8641499332F8AE9D9F0A.2_cid360?__blob=publicationFile&v=7 (abgerufen am 29.08.2019).

Corcoran et al. (2018): Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 17, Nr. 1 (2018), p. 67–75.

Deutscher Ethikrat (2017): Stellungnahme: Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung.

Finlayson et al. (2018): Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296.

Heads of Medicines Agency/European Medicines Agency (2019): HMA-EMA Joint Big Data TaskforcePhase II report: 'Evolving Data-Driven Regulation'. https://www.ema.europa.eu/en/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-evolving-data-driven-regulation_en.pdf (abgerufen am 11.02.2020).

Kim et al. (2017): Intrinsic interactive reinforcement learning – Using error-related potentials for real world human-robot interaction. Scientific reports 7, Artikelnr. 17562 (2017).

Mirsky et al. (2019): CT-GAN. Malicious Tampering of 3D Medical Imagery Using Deep Learning. arXiv 1901.03597.

Neumuth/Franke (2018): Clear oxygen-level forecasts during anaesthesia - Machine learning can predict and help interpret the risk of hypoxemia. Nature Biomedical Engineering 2, Nr. 10 (2018), p. 715–716.

Plattform Lernende Systeme (2019): Bericht der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege. https://www.plattform-lernende-systeme.de/publikationen-details/lernende-systeme-im-gesundheitswesen.html?file=files/Downloads/Publikationen/AG6_Lernende_Systeme_im_Gesundheitswesen_web_final.pdf (abgerufen am 29.08.2019).

Price (2017): Artificial Intelligence in Health Care. Applications and Legal Implications. TheSciTech Lawyer 14, Nr. 1 (2017).

PWC (2017): What doctor? Why AI and Robotics will define New Health. https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health.html (abgerufen am 29.08.2019).

The Medical Futurist (2019): FDA Approvals for Smart Algorithms in Medicine in one Giant Infographic. https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine (abgerufen am 29.08.2019).

Vial et al. (2018): The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Translational Cancer Research 7, Nr. 3 (2018), p. 803–816.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
April 2020

Bildnachweis
Tom Werner/gettyimages/Titel

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Johannes Winter (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Folgen Sie uns auf Twitter: @LernendeSysteme

Empfohlene Zitierweise
Jörn Müller-Quade et al. (Hrsg.): Sichere KI-Systeme für die Medizin – Whitepaper aus der Plattform Lernende Systeme, München 2020.