Ergebnisbericht

(gemäß Nr. 14.1 ANBest-IF)

Konsortialführung:	Universitätsklinikum Hamburg-Eppendorf
Förderkennzeichen:	01VSF18035
Akronym:	RABATT
Projekttitel:	Risikoscore für eine Algorithmenbasierte Behandlerunabhängige Aufklärung zum Therapieerfolg und zur Therapieempfehlung
Autoren:	Priv.-Doz. Dr. med. Christian-Alexander Behrendt (Projektleitung)
Förderzeitraum:	1. April 2019 – 31. März 2023

Inhaltsverzeichnis

I. Abkürzungsverzeichnis 2

II. Abbildungsverzeichnis 2

III. Tabellenverzeichnis 3

1. Zusammenfassung 4

2. Beteiligte Projektpartner 5

3. Projektziele 5

4. Projektdurchführung 7

5. Methodik 11

6. Projektergebnisse 16

7. Diskussion der Projektergebnisse 28

8. Verwendung der Ergebnisse nach Ende der Förderung 30

9. Erfolgte bzw. geplante Veröffentlichungen 31

10. Literaturverzeichnis 32

11. Anhang 35

12. Anlagen 35

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

I. Abkürzungsverzeichnis

AWMF	Arbeitsgemeinschaft der Wissenschaftlich Medizinischen Fachgesellschaften
BGB	Bürgerliches Gesetzbuch
BSG	Bundessozialgesetz
CASP	Clinical Appraisal Skills Programme
CEBM	Center for Evidence-based Medicine
DiGA	Digitale Gesundheitsanwendung
DSGVO	Datenschutzgrundverordnung
DSFA	Datenschutzfolgeabschätzung
ESVS	European Society for Vascular Surgery
EU	Europäische Union
ICD	International Classification of Diseases
KI	Künstliche Intelligenz
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Maschinelles Lernen
PAVK	Periphere arterielle Verschlusskrankheit
RABATT	Risikoscore für eine Algorithmenbasierte Behandlerunabhängige Aufklärung zum Therapieerfolg und zur Therapieempfehlung
RCT	Randomisierte kontrollierte Studie
SGB	Sozialgesetzbuch
SVS	Sicherheit in Verteilten Systemen
TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
USA	United States of America

II. Abbildungsverzeichnis

Abbildung 1: Anwendungsbeispiel für die patientenindividuelle antithrombotische Therapie auf dem Boden des thromboembolischen Gesamtrisikos (Amputation und Tod) versus schwere Blutungen im RABATT-Projekt. Bei hohem Risiko für Amputation und Tod aber geringem Blutungsrisiko ist eine aggressivere antithrombotische Therapie angezeigt, während zurückhaltende Therapiestrategien bei geringem Risiko für Amputation und Tod oder hohem Blutungsrisiko angezeigt sind. (Seite 8)

Abbildung 2: Abbildung zur Darstellung des zeitlichen Verlaufs des Projektes sowie der einzelnen methodischen Teilstudien und Zeitpunkte. (Seite 11)

Abbildung 3: Vereinfachende Taschenkarte zum GermanVasc-Summenscore zur Vorhersage von Majoramputationen und Tod bei Patienten mit symptomatischer peripherer arterieller Verschlusskrankheit (PAVK). COPD: Chronische obstruktive Lungenerkrankung. (Seite 17)

Abbildung 4: Modellkalibrierung zum Vergleich der beobachteten (blau) vs. erwarteten (grau) Risiken für Blutungsereignisse (in %) in der prospektiven GermanVasc-Kohortenstudie anhand des GermanVasc-Risikoscores bei Patienten mit Claudicatio intermittens. (Seite 18)

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Abbildung 5: Modellkalibrierung zum Vergleich der beobachteten (grün) vs. erwarteten (grau) Risiken für Blutungsereignisse (in %) in der prospektiven GermanVasc-Kohortenstudie anhand des GermanVasc-Risikoscores bei Patienten mit chronischer extremitätengefährdender Ischämie. (Seite 19)

Abbildung 6: Vereinfachende Taschenkarte zum OAC3-PAD-Summenscore zur Vorhersage von schweren Blutungsereignissen bei Patienten mit symptomatischer peripherer arterieller Verschlusskrankheit (PAVK). (Seite 20)

Abbildung 7: Modellkalibrierung zum Vergleich der beobachteten (blau) vs. erwarteten (rot) Risiken für Blutungsereignisse in der prospektiven GermanVasc-Kohortenstudie anhand des OAC3-PAD-Risikoscores. (Seite 21)

Abbildung 8: PRISMA-Flowchart der systematischen Literaturrecherche zur Assoziation zwischen Ernährung und Behandlungsergebnis bei Menschen mit peripherer arterieller Verschlusskrankheit (PAVK). (Seite 22)

III. Tabellenverzeichnis

Tabelle 1: Vergleich der zur Verfügung stehenden Variablen in Sekundärdaten der BARMER und in der prospektiven GermanVasc-Kohortenstudie. Die Operationalisierung der Variablen ist in Elixhauser et al., Kreutzberg et al., Kotov et al., Debus et al., und Peters et al. beschrieben. (Seite 13)

Tabelle 2: Basischarakteristika der Trainings- und Validierungskohorte. (Seite 18)

Tabelle 3: Basischarakteristika der Trainings- und Validierungskohorte. (Seite 20)

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

1. Zusammenfassung

Hintergrund: Für die Behandlung der peripheren arteriellen Verschlusskrankheit (PAVK) existieren zahlreiche Empfehlungen in Praxisleitlinien, die aufgrund fehlender empirischer Evidenz aus Studien auf dem Boden von Expertenmeinungen generiert wurden. Es wird derzeit diskutiert, ob die Nutzung von Versorgungsdaten die Entscheidungsfindung bei der patientenzentrierten Behandlung von Menschen mit PAVK unterstützen kann. Das vorliegende Projekt führt Beteiligte aus den Bereichen der gefäßmedizinischen Versorgungsforschung, Rechtswissenschaft und Informatik zusammen, um bestehende Versorgungsdaten für die Entwicklung von Risikovorhersagemodellen zu nutzen und dabei entstehende Fragestellungen aus den Schnittstellenbereichen zu bearbeiten.
Methodik: Es handelte sich um ein multimethodales Projekt mit datenbasierten und qualitativen Forschungsanteilen. Als Weiterverwertung der Daten und Erkenntnisse aus der IDOMENEO-Studie (01VSF16008) konnten die routinemäßig erhobenen Daten der Krankenkasse BARMER für die Entwicklung von zwei Risikovorhersagemodellen genutzt werden. Hierbei wurden maschinelle Lernverfahren verwendet und die Modelle wurden anschließend sowohl intern als auch extern mit qualitätsgesicherten Daten der prospektiven GermanVasc-Kohortenstudie (NCT03098290) validiert. Zur Anwendung der Vorhersagemodelle in der klinischen Praxis wurde eine webbasierte Schnittstelle und ein Portal zur Registrierung und Suche von präventivmedizinischen Angeboten entwickelt. Die Entwicklung und Nutzung der Risikovorhersagemodelle wurde als Anwendungsfall genutzt, um in interdisziplinären Diskussionen zwischen gefäßmedizinischen Versorgungsforschern, Rechtswissenschaftlern und Informatikern rechtliche und technische Aspekte zu identifizieren und zu bearbeiten. Während sich die Informatik primär mit privatsphärefreundlichem maschinellem Lernen und möglichen Angriffsszenarien beschäftigte, sollte die rechtswissenschaftliche Begleitung vor allem haftungs- und sozialrechtliche Themen bearbeiten.
Ergebnisse: Das RABATT-Projekt hat einen Risikoscore zur Vorhersage von Amputation und Tod nach fünf Jahren (GermanVasc-Score) für Patienten mit symptomatischer PAVK entwickelt, in den jeweils 10 prädiktive Variablen für die Subgruppe mit Claudicatio intermittens bzw. chronischer extremitätengefährdender Ischämie eingegangen sind, um die Subgruppen in jeweils fünf Risikogruppen zu diskriminieren. Der zweite Risikoscore (OAC3-PAD) konnte insgesamt acht prädiktive Variablen identifizieren, um in der gleichen Zielpopulation das Risiko für schwere Blutungen innerhalb eines Jahres vorherzusagen. Bei der Begleitung des Projektes durch die Informatik wurden verschiedene Aspekte des privatsphärefreundlichen maschinellen Lernens herausgearbeitet. Die rechtswissenschaftliche Begleitung des Projekts ergab zahlreiche haftungs- und sozialwissenschaftliche Hürden. Die Nutzung von Software mit Methoden der künstlichen Intelligenz im medizinischen Bereich stellt demnach nicht per se eine Pflichtverletzung im Haftungsrecht dar. Derartige Software hat das Potential, in Zukunft den ärztlichen Standard zu prägen. Im Umkehrschluss ist eine Nichtnutzung von derartigen Methoden und Techniken, die zum Standard geworden sind, erst nach einer Übergangsphase geeignet, um zu einer Pflichtverletzung zu führen.
Diskussion: Das RABATT-Projekt hat einen klinisch relevanten Anwendungsfall und zwei praktisch nutzbare Risikovorhersagemodelle anhand von Versorgungsdaten entwickelt. In der Anwendung beider Modelle kann die klinische Versorgungspraxis bei der Festlegung der gerinnungswirksamen Medikation unterstützt werden. In der technischen Begleitung durch die Informatik konnten zahlreiche Aspekte des privatsphärefreundlichen maschinellen Lernens und mögliche Angriffsszenarien herausgearbeitet werden. Gleichzeitig ergab die rechtswissenschaftliche Begleitung, dass die dynamische Entwicklung auf dem Gebiet der datenbasierten Forschung nur Prognosen dazu zulässt, dass diese Nutzung der Daten in Zukunft den ärztlichen Standard prägen wird, während eine haftungsrechtliche Implikation erst nach einer Übergangsphase zu erwarten ist.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

2. Beteiligte Projektpartner

***Universitätsklinikum Hamburg-Eppendorf**, Forschungsgruppe GermanVasc,*
PD Dr. med. Christian-Alexander Behrendt (Projektleitung)
Verantwortlichkeiten:
Konsortialführung und Gesamtprojektleitung

Fachlicher Ansprechpartner für Rückfragen nach Projektende:
PD Dr. med. Christian-Alexander Behrendt
Medizinisch-Wissenschaftlicher Direktor,
Deutsches Institut für Gefäßmedizinische Gesundheitsforschung gGmbH
(behrendt@hamburg.de | Telefon: +49-3212-2224422 | Telefax: +49-3212-2224422)

***Universität Hamburg**, Fakultät für Mathematik, Informatik und Naturwissenschaften,*
Arbeitsbereich Sicherheit in Verteilten Systemen (SVS),
Prof. Dr.-Ing. Hannes Federrath,
Verantwortlichkeiten: Konzeptionelle Entwicklung der Techniklösungen, Datenschutz

***BARMER**, Hauptverwaltung,*
Dr. med. Ursula Marschall (Leitende Medizinerin)
Verantwortlichkeiten:
Bereitstellung der Routinedaten über das Wissenschafts-Data-Warehouse

***Universität Hamburg**, Fakultät für Rechtswissenschaft,*
Prof. Dr. Hans-Heinrich Trute, Prof. Dr. Tilman Repgen,
Verantwortlichkeiten: Sozial- und haftungsrechtliche Begleitung

3. Projektziele

Das RABATT-Studienvorhaben verfolgte einen multimethodalen Ansatz zur Erreichung der Projektziele. Das Gesamtprojekt war in die Entwicklung geeigneter prognostischer Vorhersagemodelle, deren Validierung an externen Register- und Routinedaten, die Implementierung einer entsprechenden Techniklösung und die begleitende Evaluation und Beantwortung rechtlicher Fragestellungen (Zivilrecht, Sozialrecht) gegliedert. Die Entwicklung der Risikovorhersagemodelle erfolgte dabei durch umfassende Sekundäranalysen von Krankenkassendaten der BARMER.

Projektziel 1a und 1b: **Entwicklung eines prognostischen Vorhersagemodells**

Die heutzutage verfügbaren Praxisleitlinien enthalten zahlreiche Empfehlungen, die aufgrund mangelnder empirischer Evidenz auf dem Boden von Expertenmeinungen entwickelt wurden.[1-5] Die Anwendbarkeit dieser Leitlinienempfehlungen aber auch der Erkenntnisse aus randomisierten kontrollierten Studien ist in der klinischen Praxis im Bereich der PAVK wegen der immanenten Unterschiede zwischen Studienkohorten und der Versorgungsrealität teilweise eingeschränkt. Dies gilt insbesondere für Menschen mit multiplen Komorbiditäten und Risikofaktoren, die häufig aus rekrutierenden Studien ausgeschlossen werden. Eines der Projektziele im RABATT-Konsortialprojekt war daher die Entwicklung eines prognostischen Vorhersagemodells zur Anwendung in dem Gebiet der Behandlung von Menschen mit symptomatischer PAVK und komplexem Komorbiditätsprofil. Hierbei sollten zwei geeignete Anwendungsfälle identifiziert und die Ergebnisse und Daten der IDOMENEO-Studie (01VSF16008) weiterverwertet werden.[6, 7] Aufbauend auf den ausgewählten Anwendungsfällen sollte die Ergebnisqualität bzw. geeignete Endpunkte deskriptiv anhand von deutschen Register- und Routinedatenquellen dargestellt werden. Hieran anschließend

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

sollte eine konfirmatorische Analyse von multivariaten Modellen zur Modellentwicklung erfolgen.

Projektziel 2a und 2b: **Validierung des prognostischen Vorhersagemodells an Register- und Routinedaten**

Grundsätzlich basiert die Validität von multivariaten Modellen auf der Güte der Trainingsdaten. Die Validität von Daten aus klinischen und administrativen Registern aber auch damit entwickelte prädiktive Modelle sollten daher grundsätzlich einer regelmäßigen kontextspezifischen Überprüfung (interne und externe Validierung) unterzogen werden. Nach der Entwicklung der Vorhersagemodelle sollte deren Adjustierung und Validierung anhand von wissenschaftstheoretischen Simulationsmodellen, sowie anhand von Register- und Routinedatensätzen erfolgen. Für die Validierung von Modellen, die mit Routinedatensätzen entwickelt wurden, steht aus der IDOMENEO-Studie (01VSF16008) eine extern stichprobenartig qualitätsgesicherte Datenbasis der prospektiven GermanVasc-Kohortenstudie mit bis zu 5.608 Datensätzen und Follow-up über bis zu einem Jahr zur Verfügung.[6, 7] Bei der Validierung sind sowohl datenschutzrechtliche als auch technische Aspekte zu berücksichtigen.[8]

Projektziel 3: **Implementierung einer Big-Data-Anwendung zur Nutzung wachsender Datenbestände**

Durch den Fachbereich Informatik der Universität Hamburg sollte die Entwicklung einer Techniklösung und dessen Implementierung zur regelmäßigen Adjustierung des Vorhersagemodells erfolgen. Die Erhebung des Nutzungsverhaltens bei digitalen Gesundheitsanwendungen gehörte zu den Projektzielen in diesem Abschnitt. Gleichzeitig sollte eine browserbasierte und mobil nutzbare Schnittstelle zur Nutzung des Modells durch Patienten und das Behandlungsteam entwickelt werden. Hierbei sollten insbesondere technische Aspekte des privatsphärefreundlichen maschinellen Lernens und datenschutzrechtliche Aspekte evaluiert und bearbeitet werden. Ein weiteres Projektziel war der Aufbau und die Anbindung einer Datenbank für regionale Gefäßsportgruppen, Rauchentwöhnungs- sowie Ernährungsangebote und weitere Patienteninformationsangebote. Hierfür sollten evidenzbasierte Maßnahmen des Best Medical Treatment identifiziert und für die Nutzung in Angeboten für Patienten bereitgestellt werden. Die in diesem Projektziel entwickelten Lösungen sollten durch eine Fokusgruppe mit Experten und eine qualitative Nutzerbefragung evaluiert werden.

Projektziele 4 und 5: **Evaluation und Beantwortung zivilrechtlicher Fragestellungen bei Big-Data-Nutzung; Evaluation und Beantwortung sozialrechtlicher Fragestellungen bei Big-Data-Nutzung**

Aufgrund der zahlreichen Schnittstellen von datenbasierten Forschungsvorhaben mit Aspekten des Zivilrechts (z.B. Haftung) und Sozialrechts sollte eine kontinuierliche Begleitung der Projektschritte durch Wissenschaftler der Rechtsfakultät der Universität Hamburg erfolgen. Geplant waren Praktika, Workshops und regelmäßige Konferenzen mit allen Projektbereichen in Form von Fokusgruppendiskussionen.

Die beteiligten Rechtswissenschaftler erlangen hierdurch einen detaillierten Einblick in medizinische und versorgungsforschungswissenschaftliche Aspekte bei der Generierung und Anwendung von empirischer Evidenz in Beobachtungsstudien. Hierbei sollten relevante rechtswissenschaftliche Fragen identifiziert und am Anwendungsbeispiel bearbeitet werden. Die konkrete Bearbeitung der Fragestellungen sollte unterteilt in die Bereiche Sozialrecht und Zivilrecht erfolgen.

Für die in diesem Kapitel beschriebenen Fragestellungen und Projektziele wurden die im Kapitel 5 beschriebenen Methoden genutzt. Die Nummerierung der Gliederung entspricht sich in den Kapiteln (Projektziele, Projektdurchführung, Methodik und Ergebnisse):

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

- Projektziel 1) Entwicklung eines Vorhersagemodells zum amputationsfreien Überleben (1a) und zu Majorblutungen (1b): Retrospektive Routinedatenanalyse und Entwicklung der Vorhersagemodelle mit interner Validierung an separaten Testdaten (Seite 10).
- Projektziel 2) Validierung des Vorhersagemodells zum amputationsfreien Überleben (2a) und zu Majorblutungen (2b) mit externen Datenquellen: Validierungsstudie mit Daten aus Registern und Routinedatenquellen (Seite 11-12).
- Projektziel 3) Implementierung einer Big-Data-Anwendung zur Nutzung wachsender Datenbestände und Entwicklung einer Online-Applikation (Seite 13).
- Projektziel 4) Evaluation und Beantwortung zivilrechtlicher Fragestellungen (Seite 13).
- Projektziel 5) Evaluation und Beantwortung haftungsrechtlicher Fragestellungen (Seite 13).

4. Projektdurchführung

Das RABATT-Konsortium, bestehend aus dem Universitätsklinikum Hamburg-Eppendorf (Konsortialführung), der Fakultät für Mathematik, Informatik und Naturwissenschaften, der Fakultät für Rechtswissenschaft der Universität Hamburg sowie der gesetzlichen Krankenkasse BARMER nahm am 1. April 2019 die gemeinsame Arbeit an dem hier beschriebenen multimethodalen mehrstufigen Projekt auf.

Ausgehend von den themenrelevanten Vorarbeiten der federführend verantwortlichen Forschungsgruppe GermanVasc erfolgten verschiedene Projektschritte simultan und in einem engen Austausch mit allen projektbeteiligten Konsortialpartnern.

Zunächst erfolgte mit dem Konsortium und allen beteiligten Projektpartnern eine moderierte Fokusgruppendiskussion, in der die Einzelschritte des Projekts und die Methoden zur Erreichung der Projektziele diskutiert wurden (Projektziel 1). Diese Fokusgruppendiskussion folgte keinem a priori definierten Leitfaden sondern als offene moderierte Gruppendiskussion. Hierbei wurde über verschiedene Vorhersagemodelle und geeignete Endpunkte abgestimmt, die im klinischen Alltag von großer Relevanz sein würden. Als zentrale Endpunkte wurden die Vorhersage des amputationsfreien Überlebens (Modell 1) sowie des Blutungsrisikos (Modell 2) identifiziert, weil die hiermit eingeschlossenen Endpunkte maßgeblich für die Verschreibung der optimalen Arzneimitteltherapie (z.B. lipidsenkende und gerinnungshemmende Medikation) bei nahezu jedem Patienten mit symptomatischer PAVK sind. Beide Endpunkte wurden außerdem als sich gegenseitig ergänzende Modelle bewertet, da Betroffene mit einem hohen Risiko für den einen Endpunkt in der Regel aggressiver medikamentös behandelt werden, was sekundär zu einer erhöhten Komplikationsrate führen kann, wenn das Risiko für den zweiten Endpunkt erhöht ist. Damit ist die Abstimmung der sich ergänzenden Risiken in der klinischen Praxis gut geeignet, um die Unterstützung von prädiktiven Vorhersagemodellen zu untersuchen (**Abbildung 1**).[2, 3]

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

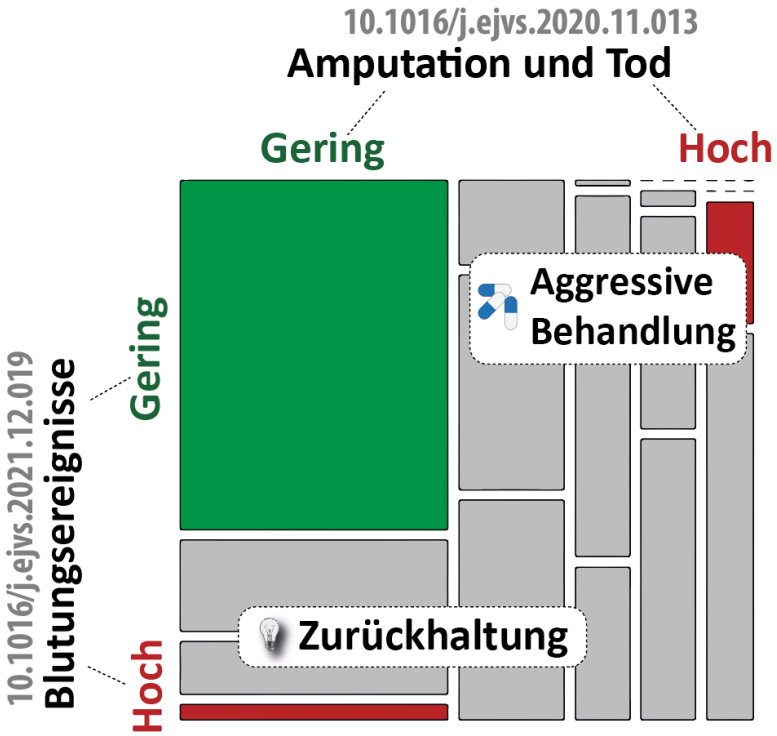

Abbildung 1: Anwendungsbeispiel für die patientenindividuelle antithrombotische Therapie auf dem Boden des thromboembolischen Gesamtrisikos (Amputation und Tod) versus schwere Blutungen im RABATT-Projekt. Bei hohem Risiko für Amputation und Tod aber geringem Blutungsrisiko ist eine aggressivere antithrombotische Therapie angezeigt, während zurückhaltende Therapiestrategien bei geringem Risiko für Amputation und Tod oder hohem Blutungsrisiko angezeigt sind.

Aufgrund der umfangreichen Vorarbeiten mit den verfügbaren Datenquellen (GermanVasc-Register und BARMER-Routinedaten) konnte im nächsten Schritt die Festlegung der geeigneten Aufgreifkriterien bzw. Variablen erfolgen.[6, 7, 9-12] Nach Festlegung des grundlegenden Studiendesigns und des statistischen Analyseplans erfolgte die Auswertung der Registerdaten der prospektiven GermanVasc-Kohortenstudie (NCT03098290) zum amputationsfreien Überleben und zu schweren Blutungskomplikationen. Anschließend erfolgte die Nutzung maschineller Lernverfahren an den Routinedaten der BARMER, wobei aufgrund der großen Anzahl an verfügbaren Variablen keine Einschränkung bzw. Einflussnahme der projektbeteiligten Forscher auf die algorithmenbasierte Variablenselektion erfolgte. In der Initialphase der Modellentwicklung wurden sowohl Verfahren des überwachten Lernens (Support Vector Machines, Nearest Neighbour) aber auch nicht-überwachten Lernens (Hidden Markov, Neurale Netzwerke) verwendet, wobei in einer Fokusgruppendiskussion mit den Fachbereichen Statistik, Informatik und Versorgungsforschung nach gewissenhafter Abwägung der Vor- und Nachteile der Klassifikations-, Regressions- und Clusterverfahren auf eine Variablenselektion mit der Least Absolute Shrinkage and Selection Operator (LASSO)-Methode zurückgegriffen wurde (Projektziel 1a). Nach der Adjustierung und internen Validierung des Risikoscores zur Vorhersage des amputationsfreien Überlebens innerhalb von 5 Jahren erfolgte die Veröffentlichung der Methode und Ergebnisse in einem fachspezifischen hochrangigen Journal und die Bereitstellung einer parallel entwickelten webbasierten Schnittstelle zur Kalkulation des individuellen Risikos.[13] Der Risikoscore wurde zudem im Konsortium und auf verschiedenen Fachkongressen vorgestellt und diskutiert und in verschiedenen externen Validierungsprojekten adressiert.

Anhand der verfügbaren qualitätsgesicherten Daten der prospektiven GermanVasc-Kohortenstudie aus dem IDOMENEO-Projekt erfolgte anschließend eine stratifizierte externe

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Validierung des Vorhersagemodells, bei dem eine moderate bis gute Diskrimination der Risikogruppen für Patientinnen und Patienten mit chronischer extremitätengefährdender Ischämie (sog. kritisches Stadium der PAVK, Fontaine-Stadien III und IV) nachgewiesen werden konnte. Die Vorhersagegüte im früheren Stadium der Claudicatio intermittens (sog. Schaufensterkrankheit, Fontaine-Stadium II) war allerdings nur moderat, was durch die Forscher unter anderem auf die im Vergleich geringe Rate an studienrelevanten Ereignissen innerhalb von einem Jahr zurückgeführt wurde (Projektziel 2a).

In der weiterführenden Projektphase erfolgte die Entwicklung des zweiten Vorhersagemodells zu schweren Blutungsereignissen innerhalb von einem Jahr nach stationärer Behandlung der symptomatischen PAVK. Hierbei konnten die Erfahrungen und Argumente aus der Entwicklung und Validierung des ersten Scores genutzt werden, um die Methoden zu verbessern. Aufgrund der guten Erfahrungen mit der Variablenselektion wurde auch hierbei die LASSO-Methode verwendet (Projektziel 1b). Zur Erreichung eines im klinischen Alltag pragmatisch anwendbaren Vorhersagesystems wurde auf dem Boden einer narrativen Literaturrecherche und Fokusgruppendiskussion mit dem Steuerkomitee der europäischen (ESVS) Leitlinien zur antithrombotischen Therapie von Gefäßkrankheiten nach der algorithmenbasierten Variablenselektion eine weitere Überprüfung und Vereinfachung der Risikofaktoren vorgenommen. Hierbei blieb die klinische Logik und Diskrimination der Risikogruppen erhalten. Der OAC3-PAD-Risikoscore beinhaltete demnach acht Variablen zur Vorhersage und wurde intern erfolgreich validiert. Die Methoden und der Risikoscore wurden fachspezifisch hochrangig publiziert und auf verschiedenen Fachkongressen vorgestellt.[14] Im Rahmen der ersten externen Validierung anhand der verfügbaren qualitätsgesicherten Studiendaten der prospektiven GermanVasc-Kohortenstudie konnte eine moderate Validität nachgewiesen und veröffentlicht werden.[15] Weitere bestätigende Validierungsprojekte wurden im Anschluss anhand französischer Routinedaten,[16] schwedischer Qualitätsregisterdaten und Registerdaten aus den USA durchgeführt (Projektziel 2b). Der OAC3-PAD-Risikoscore wurde bis zum Abschluss des Projekts in zwei aktuellen europäischen Leitlinien zur Vorhersage des Blutungsrisikos beschrieben und empfohlen.[2, 3]

Durch die projektbeteiligten Informationswissenschaftler erfolgte die Entwicklung und Bereitstellung einer Online-Schnittstelle zur responsiven Nutzung der beiden Risikoscores im klinischen Alltag (score.germanvasc.de). Die Inhalte wurden dabei in deutscher und englischer Sprache für eine Testphase bereitgestellt (Projektziel 3). Zur Bestimmung, inwiefern die projektrelevante Zielpopulation digitale Gesundheitsanwendungen nutzt, wurde außerdem eine multizentrische Befragung von Patientinnen und Patienten durchgeführt (Projektziel 3).[17] Im Rahmen der Pilottestung der webbasierten Anwendung mit etwa 600 Nutzern der wurden auch verschiedene Befragungen der Patient:innen durchgeführt, um die Erreichbarkeit dieser Zielpopulation durch die verschiedenen digitalen Technologien zu evaluieren. Im Rahmen einer repräsentativen Querschnitts Survey Studie konnten 326 Patient:innen mit stationären Behandlungen der symptomatischen PAVK an 13 spezialisierten deutschen Einrichtungen konsekutiv befragt werden. Hierbei konnten insbesondere auch die Verfügbarkeit und Nutzung von Smartphones, mobilen Gesundheitsanwendungen (DiGA) und sogenannten Wearables bestimmt werden (Projektziel 3).[17]

Nach initialen Verzögerungen bei der Rekrutierung geeigneter Promotionswissenschaftler in der Fakultät für Rechtswissenschaft, die durch die beiden projektleitenden Wissenschaftler der Fakultät für Rechtswissenschaft überbrückt wurden, konnten die beiden Promovenden während der Validierungsphase der Scores ihre Arbeit aufnehmen. Aufgrund der globalen Pandemie und den einsetzenden Einschränkungen in der ersten Jahreshälfte 2020 mussten jedoch die Vor-Ort-Begleitungen in Form von onlinebasierten Diskussionen via Zoom oder Teams stattfinden. Demnach fanden zwischen 2020 und 2022 zahlreiche Online-Fokusgruppendiskussionen und Workshops statt, bei denen den Rechtswissenschaftlern fundierte Einblicke in die Generierung von Evidenz, deren Implementierung in

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Leitlinienempfehlungen sowie die Anwendung in der klinischen Praxis gewährt wurde (Projektziel 4 und 5).

Parallel zu den vorbeschriebenen Projektschritten erfolgte eine Fokusgruppendiskussion mit dem Konsortium und die umfangreiche narrative Literaturrecherche sowie Analyse der verfügbaren Praxisleitlinien, um evidenzbasierte Empfehlungen zum sogenannten Best Medical Treatment in der Behandlung von Menschen mit PAVK zu identifizieren (Projektziel 3). Die Projektleitung war außerdem an der Entwicklung und Überarbeitung verschiedener nationaler und internationaler Praxisleitlinien zur PAVK beteiligt, so dass Zugriff auf die systematischen Literaturrecherchen der Leitliniensteuerkomitees bestand. Da nur wenige evidenzbasierte Empfehlungen zum Thema der Ernährung in den Praxisleitlinien existierten, führte die Projektleitung eine systematische Literaturrecherche und eine Patientenbefragung hierzu durch (Projektziel 3).[18, 19] Alle Erkenntnisse und Empfehlungen wurden anschließend im Rahmen einer Diskussion mit der Patientenvertretung abgestimmt und in einem laienverständlichen Informationsartikel lizenzfrei veröffentlicht.[20]

Im Fachbereich Informatik der Universität Hamburg wurde begleitend der Themenkomplex des sogenannten privatsphäre-freundlichen maschinellen Lernens sowie Gegenmaßnahmen gegen Privatsphäreangriffe bearbeitet, was in verschiedenen Publikationen und einem Thesenpapier mündete (Projektziel 3, 4 und 5).[21, 22] Hierbei erfolgte eine enge Abstimmung mit den Bereichen gefäßmedizinische Versorgungsforschung und Rechtswissenschaft, um die theoretischen Aspekte mit geeigneten Anwendungsfällen abzugleichen.

Aufgrund von Problemen bei der Personalbeschaffung und mit der COVID-19-Pandemie assoziierten Ursachen, insbesondere bei der rechtswissenschaftlichen Begleitung und im Fachbereich Informatik des Konsortialpartners Universität Hamburg, hat die Projektleitung im März 2023 eine ausgabenneutrale Laufzeitverlängerung für das RABATT-Projekt beantragt. Hierdurch wurde das Projekt über insgesamt 48 Monate durchgeführt (**Abbildung 2**). Aufgrund von Verzögerungen bei der Bearbeitung des haftungsrechtlichen Projektanteils der Rechtswissenschaftler und dem Abbruch der initial geplanten Promotionsarbeit erfolgte ein Antrag auf ausgabenneutrale Umwidmung der Mittel und Änderung des Arbeitsplans, dem am 16. Januar 2023 zugestimmt wurde. Demnach konnte für den haftungsrechtlichen Teilprojektanteil unter der Supervision des zuständigen Teilprojektleiters ein Honorargutachten angefertigt werden.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

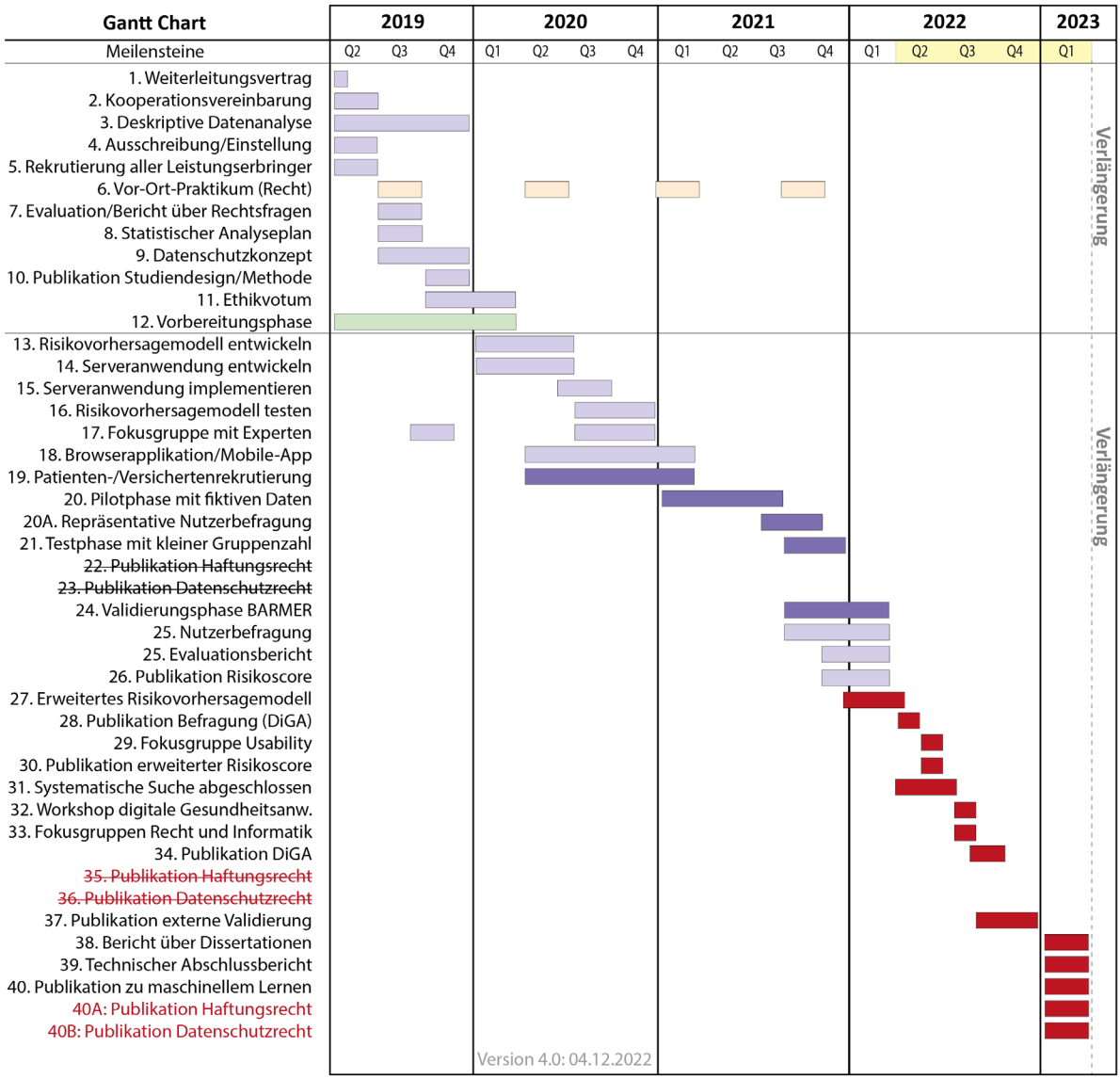

Abbildung 2: Abbildung zur Darstellung des zeitlichen Verlaufs des Projektes sowie der einzelnen methodischen Teilstudien und Zeitpunkte.

5. Methodik

Projektziel 1a: **Entwicklung eines Vorhersagemodells zum amputationsfreien Überleben**

Volltext im Anhang:

Kreutzburg T, Peters F, Kuchenbecker J, Marschall U, Lee R, Kriston L, Debus ES, Behrendt CA. Editor's Choice - The GermanVasc Score: A Pragmatic Risk Score Predicts Five Year Amputation Free Survival in Patients with Peripheral Arterial Occlusive Disease. Eur J Vasc Endovasc Surg. 2021 Feb;61(2):248-256. doi: 10.1016/j.ejvs.2020.11.013. Epub 2020 Dec 15. PMID: 33334671.

Es wurde eine retrospektive Analyse von routinemäßig erhobenen faktisch anonymisierten administrativen Daten (Routinedaten) der BARMER mit ca. 9 Millionen Versicherten durchgeführt. Faktisch anonymisiert meint, dass die notwendigen technischen und administrativen Maßnahmen getroffen wurden, um eine Re-Identifizierung nur mit einem unverhältnismäßigem Aufwand zu ermöglichen. Die zur Verfügung stehenden Abrechnungsdaten beschreiben die sektorenübergreifende medizinische Versorgung der Patienten, sofern diese gegenüber dem Kostenträger in Rechnung gestellt wurden. Für die Identifizierung der geeigneten Aufgreifkriterien wurde die deutsche Modifikation der

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

International Classification of Diseases (ICD) in ihrer 10. Revision genutzt. Die entsprechenden Klassifikationssystematiken für Diagnosen, Prozeduren und Arzneimittelverordnungen orientierten sich an den einschlägigen methodischen Empfehlungen, z.B. Elixhauser et al. und Quan et al..[23, 24] Die Gesamtkohorte wurde a priori in eine Trainings- vs. Validierungskohorte unterteilt (60:40%). Dieser Wert wurde auf dem Boden einer narrativen Literaturanalyse gewählt.

Ein- und Ausschlusskriterien: Versicherte mit einem Alter ab 40 Jahren mit stationären Indexbehandlungen in deutschen Krankenhäusern mit einer Hauptdiagnose PAVK im Stadium II bis IV nach Fontaine zwischen 1. Januar 2008 und 31. Dezember 2016. Für die Identifizierung der Indexbehandlung wurde eine Rückschauzeit von 3 Jahren genutzt. Patienten mit vorheriger Majoramputation oder Tod innerhalb von 30 Tagen nach Entlassung wurden ausgeschlossen. Einen Ausschluss nach Versicherungszeiten gab es nicht.

Die Variablen beinhalteten Alter, Geschlecht, Rauchen, 30 verschiedene Elixhauser Komorbiditätsgruppen, Jahr der Entlassung, vorhergehender Herzinfarkt oder Schlaganfall, Vorhofflimmern, Dialyse, Gangrän (PAVK IV mit Gangrän), Entlassungsziel Rehabilitation oder Pflegeeinrichtung, Hospizverlegung, 190 Nebendiagnosen bei Aufnahme. Die Gruppierung erfolgte in Altersgruppen. Fehlende Daten (0,5%) wurden durch fallweisen Ausschluss berücksichtigt.

Der primäre Endpunkt umfasste jede Majoramputation oberhalb des Knöchels oder Tod innerhalb von fünf Jahren.

Die Modellentwicklung erfolgte in 9 Schritten: Stratifizierung, Trennung in Trainings- und Validierungskohorte, algorithmenbasierte Variablenselektion (LASSO) mit Strafterm, Breiman Permutation zur Identifizierung und Sortierung der Top-10-Variablen nach dem Impact, Anpassung der Cox-Regression, interne Modellvalidierung (Diskrimination), Kalibrierung des Modells, Erstellung der Risikogruppen mit Kaplan-Meier-Funktionen, Erstellung des zusammenfassenden Score-Sheets. Als Sensitivitätsanalyse erfolgte ein Elastic Net Approach.

Projektziel 2a: **Externe Validierung des Vorhersagemodells zum amputationsfreien Überleben**

Es wurde eine retrospektive Analyse von prospektiv erhobenen und qualitätsgesicherten Studiendaten aus der GermanVasc-Kohortenstudie (NCT03098290) durchgeführt. Die verfügbaren Studiendaten beinhalteten Patienten mit stationärer invasiver Behandlung der symptomatischen PAVK (Stadium II bis IV nach Fontaine) zwischen 1. Mai 2018 und 31. Dezember 2021 an 37 deutschen Gefäßzentren. Das Follow-up schloss den Zeitraum bis März 2022 ein und beinhaltete die relevanten Endpunkte (Majoramputation oder Tod) nach 3, 6 und 12 Monaten nach Krankenhausentlassung. Zur Verfügung standen 16 Variablen, außer Alkoholmissbrauch, Elektrolytstörung und Demenz.

Fehlende Daten wurden durch Imputationsverfahren berücksichtigt. Die Zeit von der Behandlung bis zum Auftreten des Events wurde zensiert nach einem Jahr für die Cox-Regression. Zur Beurteilung der Modellgüte bzw. Diskrimination der Risikogruppen erfolgte die Berechnung von Harrell´s C (Concordance Index) und der Vergleich der 95%-Konfidenzintervalle (beobachtet vs. erwartet).[25] Hierdurch kann eine objektivierbare Beurteilung erfolgen, wobei kein Konsens über Grenzwerte für eine moderate oder gute Modellgüte bzw. Diskrimination existiert. In der Regel wird ein C-Wert unter 0,5 als unzureichend und über 0,6-0,8 als moderat beschrieben. Die Berichterstattung folgte dem Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)-Statement.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Tabelle 1: Vergleich der zur Verfügung stehenden Variablen in Sekundärdaten der BARMER und in der prospektiven GermanVasc-Kohortenstudie. Die Operationalisierung der Variablen ist in Elixhauser et al., Kreutzberg et al., Kotov et al., Debus et al., und Peters et al. beschrieben.

Variable	In BARMER-Daten	In GermanVasc-Daten
Alter	Ja (in Jahren)	Ja (in Jahren)
Geschlecht	Ja (dichotomisiert)	Ja (M/W/D)
Suchtstörungen, Rauchen	Ja (Elixhauser)	Ja
Diabetes	Ja (Elixhauser)	Ja
Dialysepflichtigkeit	Ja (Elixhauser)	Ja
Alkoholabhängigkeit	Ja (Elixhauser)	Nein
Elektrolytstörungen	Ja (Elixhauser)	Nein
Krebs	Ja (Elixhauser)	Ja
COPD	Ja (Elixhauser)	Ja
Fettstoffwechselstörung	Ja (Elixhauser)	Ja
Vorhergehende Krankenhausaufenthalte	Ja	Ja
Demenz	Ja (Elixhauser)	Nein
Ulcera/Nekrosen (bei PAVK)	Ja (Elixhauser)	Ja
Herzinsuffizienz	Ja (Elixhauser)	Ja
Kardiale Arrhythmien	Ja (Elixhauser)	Ja
Niereninsuffizienz	Ja (Elixhauser)	Ja

Projektziel 1b: **Entwicklung eines Vorhersagemodells zum Auftreten von schweren Blutungskomplikationen**

Volltext im Anhang:

Behrendt CA, Kreutzburg T, Nordanstig J, Twine CP, Marschall U, Kakkos S, Aboyans V, Peters F. The OAC3-PAD Risk Score Predicts Major Bleeding Events one Year after Hospitalisation for Peripheral Artery Disease. Eur J Vasc Endovasc Surg. 2022 Mar;63(3):503-510. doi: 10.1016/j.ejvs.2021.12.019. Epub 2022 Feb 4. PMID: 35125278.

Es wurde eine retrospektive Analyse von routinemäßig erhobenen faktisch anonymisierten administrativen Daten (Routinedaten) der BARMER mit ca. 9 Millionen Versicherten durchgeführt. Die zur Verfügung stehenden Abrechnungsdaten beschreiben die sektorenübergreifende medizinische Versorgung der Patienten, sofern diese gegenüber dem Kostenträger in Rechnung gestellt wurden. Für die Identifizierung der geeigneten Aufgreifkriterien wurde die deutsche Modifikation der International Classification of Diseases (ICD) in ihrer 10. Revision genutzt. Die entsprechenden Klassifikationssystematiken für Diagnosen, Prozeduren und Arzneimittelverordnungen orientierten sich an den einschlägigen methodischen Empfehlungen, z.B. Elixhauser et al. und Quan et al..[23, 24]. Die Gesamtkohorte wurde a priori in eine Trainings- vs. Validierungskohorte unterteilt (60:40%).

Ein- und Ausschlusskriterien: Versicherte mit einem Alter ab 40 Jahren mit stationären Indexbehandlungen in deutschen Krankenhäusern mit einer Hauptdiagnose PAVK im Stadium II bis IV nach Fontaine zwischen 1. Januar 2010 und 31. Dezember 2018. Für die Identifizierung der Indexbehandlung wurde eine Rückschauzeit bis 1. Januar 2005 genutzt. Patienten, die

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

innerhalb des Krankenhaufenthaltes verstarben oder deren Aufenthalt länger als 100 Tage dauerte, wurden ausgeschlossen. Einen Ausschluss nach Versicherungszeiten gab es nicht.

Für die Entwicklung der Leitlinien der European Society for Vascular Surgery (ESVS) zur antithrombotischen Therapie von Gefäßkrankheiten wurden systematische Literaturübersichten erstellt. Die klinische Relevanz der Variablen wurde vom involvierten Leitlinienkomitee diskutiert. Die Variablen beinhalteten Alter, Geschlecht, Rauchen, Alkohol- oder Drogenmissbrauch, Herzinsuffizienz, Bluthochdruck, Lebererkrankungen, vorhergehende transiente ischämische Attacke oder Schlaganfall, Anämie, Demenz und schwere Niereninsuffizienz. Vorhergehende Blutungsereignisse wurden durch folgende Kodierungen identifiziert: Transfusion, Koagulopathie, Hauptdiagnose einer schweren Blutung während des vorhergehenden Jahres.

Der primäre Endpunkt Majorblutung nach einem Jahr orientierte sich an den adaptierten Kriterien der International Society on Thrombosis and Haemostasis (ISTH).

Die Modellentwicklung erfolgte in 9 Schritten: Trennung in Trainings- und Validierungskohorte, algorithmenbasierte Variablenselektion (LASSO) mit Strafterm, Anpassung der Cox-Regression, Erstellung der Risikogruppen mit Kaplan-Meier-Funktionen, interne Modellvalidierung (Diskrimination), Kalibrierung des Modells, Erstellung des zusammenfassenden Score-Sheets. Als Sensitivitätsanalyse erfolgte eine Berechnung der Endpunkte nach drei Jahren, Ausschluss von Patienten mit Antithrombotika sowie eine geschlechterstratifizierte Analyse.

Projektziel 2b: **Externe Validierung des Vorhersagemodells zum Auftreten von schweren Blutungskomplikationen**

Volltext im Anhang:

Peters F, Behrendt CA. External Validation of the OAC3-PAD Risk Score to Predict Major Bleeding Events Using the Prospective GermanVasc Cohort Study. Eur J Vasc Endovasc Surg. 2022 Oct;64(4):429-430. doi: 10.1016/j.ejvs.2022.07.055. Epub 2022 Aug 8. PMID: 35952908.

Es wurde eine retrospektive Analyse von prospektiv erhobenen und qualitätsgesicherten Studiendaten aus der GermanVasc-Kohortenstudie (NCT03098290) durchgeführt. Die verfügbaren Studiendaten beinhalteten Patienten mit stationärer invasiver Behandlung der symptomatischen PAVK (Stadium II bis IV nach Fontaine) zwischen 1. Mai 2018 und 31. Dezember 2021 an 37 deutschen Gefäßzentren.[6] Das Follow-up schloss den Zeitraum bis März 2022 ein und beinhaltete die relevanten Endpunkte (schwere Blutungsereignisse) nach 3, 6 und 12 Monaten nach Krankenhausentlassung. Zur Verfügung standen 16 Variablen, außer Anämie und Demenz (Tabelle 1). Die letztgenannte Variable wurde durch den Pflege- und Mobilitätsstatus (Bettlägerig) approximiert. Die Variable Dyslipidämie wurde durch Statintherapie approximiert. Fehlende Daten wurden durch Imputationsverfahren berücksichtigt. Die Zeit von der Behandlung bis zum Auftreten des Events wurde zensiert nach einem Jahr für die Cox-Regression. Zur Beurteilung der Modellgüte bzw. Diskrimination der Risikogruppen erfolgte die Berechnung von Harrell´s C und der Vergleich der 95%-Konfidenzintervalle (beobachtet vs. erwartet). Die Berichterstattung folgte dem TRIPOD-Statement.

Projektziel 3: **Befragung zur Nutzung digitaler Gesundheitsanwendungen**

Volltext im Anhang:

Alushi K, Hinterseher I, Peters F, Rother U, Bischoff MS, Mylonas S, Grambow E, Gombert A, Busch A, Gray D, Konstantinou N, Stavroulakis K, Horn M, Görtz H, Uhl C, Federrath H, Trute HH, Kreutzburg T, Behrendt CA. Distribution of Mobile Health Applications amongst Patients with Symptomatic Peripheral Arterial Disease in Germany: A Cross-Sectional Survey Study. J Clin Med. 2022 Jan 19;11(3):498. doi: 10.3390/jcm11030498. PMID: 35159950; PMCID: PMC8836389.

Eine Querschnittsbefragung (qualitative Nutzerbefragung) von allen konsekutiv stationär behandelten Patienten mit symptomatischer PAVK an 12 universitären und einer nicht-

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

universitären Einrichtung wurde durchgeführt, um das Nutzungsverhalten bei digitalen Gesundheitsanwendungen zu erheben und die Akzeptanz einer webbasierten Schnittstelle zu bestimmen. Der Erhebungszeitraum umfasste einen flexiblen Zeitraum von insgesamt 30 zusammenhängenden Tagen in den teilnehmenden Zentren, wobei ein Startdatum zwischen 1. Juli 2021 und 1. September 2021 gewählt werden konnte. Nur Patienten mit stationärer Behandlung der Claudicatio intermittens (PAVK im Stadium IIA oder IIB nach Fontaine) oder chronischen extremitätengefährdenden Ischämie (PAVK im Stadium III oder IV nach Fontaine) wurden befragt und eingeschlossen. Der Fragebogen (Anhang) wurde im Rahmen einer Gruppendiskussion mit allen projektbeteiligten Wissenschaftler:innen und den freiwillig teilnehmenden Gefäßmediziner:innen bei einer Online-Konferenz sowie in mehreren Überarbeitungsrunden via E-Mail entwickelt und papierbasiert in den teilnehmenden Zentren zur Verfügung gestellt und zentral verarbeitet. Es wurden faktisch anonymisierte Daten erhoben. Die erhobenen Daten wurden deskriptiv analysiert und zur Identifizierung von Faktoren, die mit der Nutzung von digitalen Gesundheitsanwendungen assoziiert waren, durch eine Rückwärts-Variablenselektion in die multivariaten Modelle eingefügt. In die multivariaten Modelle gingen Alter, Geschlecht, Bildungsabschluss, Siedlungsdichte des Wohnraums und Stadium der PAVK ein (Alushi et al. 2021). Die gleiche Fokusgruppe hat sich in zwei weiteren Online-Konferenzen moderiert zu der webbasierten Schnittstelle ausgetauscht und die von den Informationswissenschaftlern bzw. Entwicklern vorgestellten Entwürfe kommentiert (siehe nachfolgende Methodik zu Projektziel Fokusgruppendiskussionen).

Die durch die Softwareentwickler und Informationswissenschaftler entwickelte webbasierte Schnittstelle zur Risikovorhersage (https://score.germanvasc.de) wurde mit den in der Nutzerbefragung eingeschlossenen Patient:innen und in den teilnehmenden Zentren tätigen ärztlichen und pflegerischen Mitarbeiter:innen zum Zeitpunkt der erstmaligen Onlinestellung getestet. Hierfür wurden die Teilnehmenden über den Start informiert und gebeten, die Funktionen auszuprobieren, bevor die Fragebögen ausgegeben wurden.

Projektziel 3: **Systematische Literaturrecherche zum Einfluss der Ernährung auf Behandlungsergebnisse**

Volltext im Anhang:
Adegbola A, Behrendt CA, Zyriax BC, Windler E, Kreutzburg T. The impact of nutrition on the development and progression of peripheral artery disease: A systematic review. Clin Nutr. 2022 Jan;41(1):49-70. doi: 10.1016/j.clnu.2021.11.005. Epub 2021 Nov 11. PMID: 34864455.

Eine systematische Literaturrecherche der Datenbanken bei PubMed wurde durchgeführt zu den Suchbegriffen „PAVK“ und „Ernährung“ mit entsprechenden Synonymen und in englischer Sprache. Randomisierte kontrollierte Studien, Beobachtungsstudien, Kohortenstudien, Fall-Kontroll-Studien und Querschnittsstudien zum Zusammenhang zwischen PAVK und Ernährung mit Publikationsdatum zwischen Januar 1974 und Dezember 2019 wurden berücksichtigt und gescreent (P: Patienten mit PAVK, I: Strukturierte Ernährungsmaßnahmen und Diäten, C: Studienspezifischer Standard, O: Inzidenz einer PAVK bei gesunden Probanden oder Outcome der PAVK-Behandlung bei betroffenen Patienten). Das Protokoll zur Suche wurde a priori bei PROSPERO (CRD42020204398) registriert. Zwei Autoren führten die Suche durch. Die Berichterstattung folgte dem PRISMA-Statement. Eine Qualitätsevaluation der eingeschlossenen Studien erfolgte mit dem Cochrane Collaboration Risk of Bias Tool (RCT) bzw. Critical Appraisal Skills Programme (CASP) oder Center for Evidence-based Medicine (CEBM) für Beobachtungsstudien.

Projektziel 3 & 4 & 5: **Fokusgruppendiskussionen**

Im RABATT-Projekt wurden insgesamt fünf moderierte Fokusgruppendiskussionen durchgeführt, um Impulse und Gruppenmeinungen zu den folgenden Themen zu generieren. Einen strukturierten Leitfaden gab es a priori nicht. Die Moderation der Diskussionen erfolgte durch die wissenschaftliche Projektleitung:

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

a) Geeignete Anwendungsfälle und Endpunkte für die beiden Risikovorhersagemodelle,

b) Aufbau und Funktionalität von webbasierten Schnittstellen zur Risikovorhersage,

c) Relevante rechtswissenschaftliche Fragestellungen aus dem Bereich Zivilrecht,

d) Relevante rechtswissenschaftliche Fragestellungen aus dem Bereich Sozialrecht,

e) Privatsphärefreundliches maschinelles Lernen und Angriffsszenarien in der klinischen Anwendung,

f) Umfang von evidenzbasierten Empfehlungen zum sogenannten Best Medical Treatment bei Patienten mit PAVK.

Jede der vorgenannten Fokusgruppendiskussionen erfolgte unter Einbindung der projektbeteiligten Forscher:innen und Fachbereiche, wobei eine inklusive Einbeziehung von Dritten gestattet und gewünscht war. Mindestens beteiligt waren jeweils die wissenschaftliche Projektleitung, Expert:innen aus dem Bereich Epidemiologie und Statistik der Forschungsgruppe GermanVasc, die Projektleiter der Konsortialpartner und deren Promovenden sowie interessierte Forscher:innen aus den Bereichen mit Fachbezug. Für die Diskussion zu b) Aufbau und Funktionalität von webbasierten Schnittstellen zur Risikovorhersage wurden die an der Nutzerbefragung beteiligten Zentrumsleiter eingeladen. Die homogene Auswahl der Teilnehmer erfolgte unter der Zielvorgabe, möglichst gleichermaßen betroffene Personen einzubinden. Es erfolgte im Rahmen der Fokusgruppendiskussionen jeweils eine Einleitung zum Thema mit Präsentation der zentralen Fragestellung und anschließend eine moderierte Diskussion.

Bei der Nutzerbefragung erfolgte der Einschluss konsekutiv über 30 Tage behandelter Patient:innen an interessierten Zentren (siehe Projektziel 3).

6. Projektergebnisse

Das RABATT-Projekt hat zahlreiche Ergebnisse zu den einzelnen Fragestellungen ergeben. Die Methoden und Ergebnisse der einzelnen Studien bzw. qualitativen Forschungsinhalte sind in den angefügten Publikationen im Detail erläutert. In diesem Bericht wird jeweils auf die Veröffentlichung verwiesen (jeweils als Volltext/PDF im Anhang verfügbar).

Projektziel 1a: **Entwicklung eines Vorhersagemodells zum amputationsfreien Überleben**

Volltext im Anhang:

Kreutzburg T, Peters F, Kuchenbecker J, Marschall U, Lee R, Kriston L, Debus ES, Behrendt CA. Editor's Choice - The GermanVasc Score: A Pragmatic Risk Score Predicts Five Year Amputation Free Survival in Patients with Peripheral Arterial Occlusive Disease. Eur J Vasc Endovasc Surg. 2021 Feb;61(2):248-256. doi: 10.1016/j.ejvs.2020.11.013. Epub 2020 Dec 15. PMID: 33334671.

Insgesamt sind 87.293 Patienten mit Behandlung der symptomatischen PAVK (45,3% Frauen, mittleres Alter 71,4 Jahre ± 11,1 Jahre) zwischen 1. Januar 2008 und 31. Dezember 2016 in die Analysen eingegangen (ca. 10.000 pro Jahr). Die Basischarakteristika der Trainings- und Validierungskohorte sind in Tabelle 2 dargestellt. Hiervon wurden 46.703 mit einer Claudicatio intermittens und 40.590 mit einer chronischen extremitätengefährdenden Ischämie behandelt. Innerhalb von fünf Jahren nach Krankenhausentlassung sind 19% der Patienten mit Claudicatio intermittens und 50% mit chronischer extremitätengefährdender Ischämie verstorben. Jeweils 1% bzw. 9% wurden in diesem Zeitpunkt amputiert und der Kombinationsendpunkt trat in 20% bzw. 52% auf.

Für die Gruppe an Patienten mit Claudicatio intermittens wurden höheres Alter, Dialysepflichtigkeit, Alkoholabhängigkeit, Elektrolytstörungen, Krebs, Diabetes, chronisch obstruktive Lungenerkrankung, männliches Geschlecht, keine Dyslipidämie und vorhergehende Krankenhausbehandlungen als Risikofaktoren identifiziert. Diese Variablen und deren Gewichtung ergaben eine gute Diskrimination der fünf Risikogruppen (c = 0,70,

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

95%-Konfidenzintervall 0,69-0,71) (**Abbildung 3**).Für die Gruppe an Patienten mit chronischer extremitätengefährdender Ischämie wurden hohes Alter, vaskuläre bzw. unspezifische Demenz, Dialysepflichtigkeit, Gangrän, Krebs, Herzinsuffizienz, Elektrolytstörungen, Niereninsuffizienz und kardiale Arrhythmien als Risikofaktoren identifiziert und die Modellgüte war gleichermaßen gut (c = 0,69, 95%-Konfidenzintervall 0,67-0,71) (**Abbildung 3**).

Die Kalibrierung ergab gute Übereinstimmungen zwischen beobachteten und erwarteten Werten und die Sensitivitätsanalysen waren bestätigend. Weitere Details finden sich im Anhang (Kreutzburg T, et al. 2021).

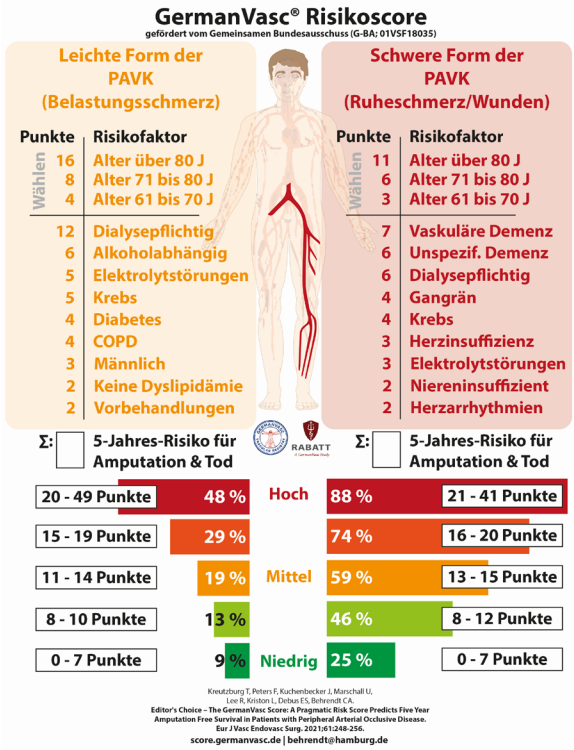

Abbildung 3: Vereinfachende Taschenkarte zum GermanVasc-Summenscore zur Vorhersage von Majoramputationen und Tod bei Patienten mit symptomatischer peripherer arterieller Verschlusskrankheit (PAVK). COPD: Chronische obstruktive Lungenerkrankung.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Tabelle 2: Basischarakteristika der Trainings- und Validierungskohorte.

	Total N= 87,293	IC Training N=28,021	IC Validation N=18,682	CLTI Training N=24,354	CLTI Validation N=16,236
Mittleres Alter in Jahren +- SD	71.4±11.1	68.8±10.1	69.0±10.2	74.5±11.3	74.4±11.3
Weibliches Geschlecht (%)	39,545 (45.3)	12,046 (43.0)	8050 (43.1)	11,684 (48.0)	7765 (47.8)
Elixhauser Gruppen (3 Jahre)					
Herzinsuffizienz (%)	20,399 (23.4)	4081 (14.6)	2698 (14.4)	8200 (33.7)	5420 (33.4)
Kardiale Arrhythmien (%)	21,351 (24.5)	4530 (16.2)	3081 (16.5)	8205 (33.7)	5535 (34.1)
Hypertension (%)	68,384 (78.3)	21,437 (76.5)	14,316 (76.6)	19,587 (80.4)	13,044 (80.3)
Diabetes, kompliziert (%)	24,004 (27.5)	4116 (14.7)	2792 (14.9)	10,268 (42.2)	6828 (42.1)
Niereninsuffizienz (%)	23,728 (27.2)	5045 (18.0)	3533 (18.9)	9074 (37.3)	6076 (37.4)
Chronische Lungenerkrankungen (%)	12,634 (14.5)	3655 (13.0)	2507 (13.4)	3915 (16.1)	2557 (15.7)
Übergewicht (%)	11,656 (13.4)	3237 (11.6)	2073 (11.1)	3813 (15.7)	2533 (15.6)
Psychische und Verhaltensstörungen durch Tabak (%)	15,330 (17.6)	5919 (21.1)	3880 (20.8)	3385 (13.9)	2146 (13.2)
Vorhergehender Herzinfarkt (%)	7465 (8.6)	2218 (7.9)	1499 (8.0)	2274 (9.3)	1474 (9.1)
Vorhergehender Schlaganfall (%)	7307 (8.4)	1545 (5.5)	1031 (5.5)	2846 (11.7)	1885 (11.6)
Vorhofflimmern (%)	12,801 (14.7)	2379 (8.5)	1580 (8.5)	5303 (21.8)	3539 (21.8)
Entlassung in Pflegeeinrichtung (%)	2076 (2.4)	119 (0.4)	86 (0.5)	1141 (4.7)	730 (4.5)
Entlassung in Rehabilitation (%)	5367 (6.1)	1147 (4.1)	761 (4.1)	2109 (8.7)	1350 (8.3)
Dialysepflichtigkeit (%)	1812 (2.1)	180 (0.6)	122 (0.7)	880 (3.6)	630 (3.9)
Diagnose (Indexaufenthalt)					
Dyslipidämie (E78, %)	28,580 (32.7)	10,812 (38.6)	7111 (38.1)	6399 (26.3)	4258 (26.2)
Demenz (F03, %)	1711 (2.0)	107 (0.4)	73 (0.4)	918 (3.8)	613 (3.8)
Mediane Anzahl Medikamente [IQR]	9 [5, 13]	8 [5, 11]	8 [5, 12]	11 [7, 16]	11 [7, 16]
Medianes Follow-up [IQR]	1503 [83, 1825]	1825 [1134, 1825]	1804 [1127, 1825]	1131 [481, 1825]	1144 [475, 1825]
Antithrombotika (%)	40,437 (46.3)	11,760 (42.0)	8018 (42.9)	12,353 (50.7)	8306 (51.2)
Lipidsenker (%)	39,431 (45.2)	14,271 (50.9)	9546 (51.1)	9280 (38.1)	6334 (39.0)
Antihypertensiva (%)	72,602 (83.2)	22,557 (80.5)	15,041 (80.5)	20,971 (86.1)	14,033 (86.4)
Tod oder Amputation (%)	30,635 (35.1)	5514 (19.7)	3728 (20.0)	12,852 (52.8)	8541 (52.6)
Gesamtsterblichkeit (%)	29,129 (33.4)	5323 (19.0)	3592 (19.2)	12,118 (49.8)	8096 (49.9)
Majoramputation (%)	4256 (4.9)	401 (1.4)	255 (1.4)	2169 (8.9)	1431 (8.8)
Myokardinfarkt (%)	10,180 (11.7)	3073 (11.0)	2142 (11.5)	3007 (12.3)	1958 (12.1)
Schlaganfall (%)	11,874 (13.6)	3541 (12.6)	2295 (12.3)	3596 (14.8)	2442 (15.0)

Projektziel 2a: **Externe Validierung des Vorhersagemodells zum amputationsfreien Überleben**

Zwischen Mai 2018 und Dezember 2021 wurden insgesamt 5.479 Patienten mit symptomatischer PAVK an 37 Gefäßzentren invasiv behandelt, zu denen vollständige Follow-up-Daten vorlagen (33,1% Frauen, mittleres Alter 69,0 Jahre, 67,6% Claudicatio intermittens, 33,0% offen-chirurgische Revaskularisation). Die Modelldiskrimination war moderat für die Subgruppe an Patienten mit Claudicatio intermittens (c = 0,66, 95%-Konfidenzintervall 0,58-0,74, **Abbildung 4**) und gut für Patienten mit chronischer extremitätengefährdender Ischämie (c = 0,69, 95%-Konfidenzintervall 0,65-0,73, **Abbildung 5**).

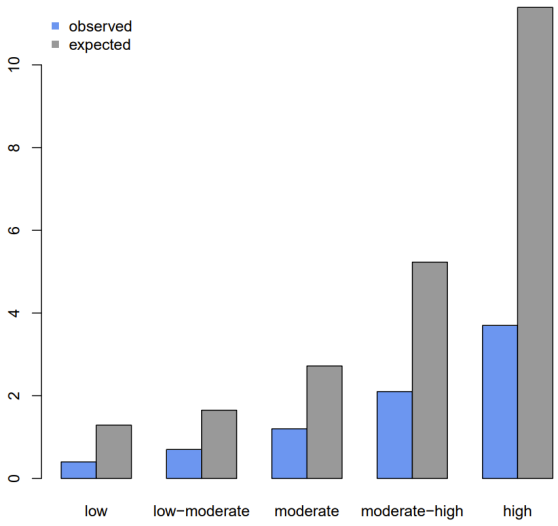

Abbildung 4: Modellkalibrierung zum Vergleich der beobachteten (blau) vs. erwarteten (grau) Risiken für Blutungsereignisse (in %) in der prospektiven GermanVasc-Kohortenstudie anhand des GermanVasc-Risikoscores bei Patienten mit Claudicatio intermittens.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

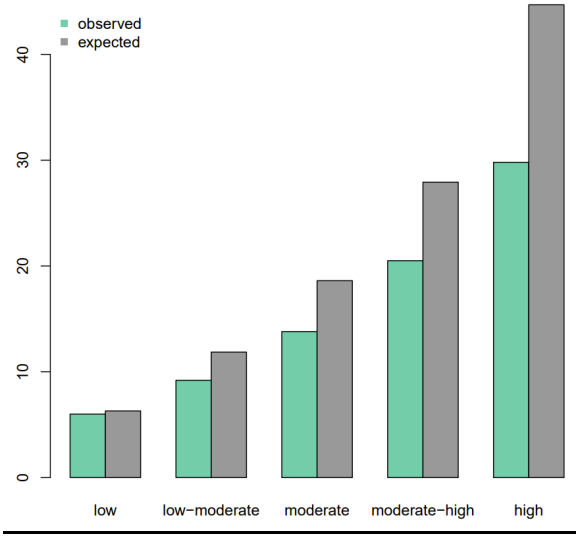

Abbildung 5: Modellkalibrierung zum Vergleich der beobachteten (grün) vs. erwarteten (grau) Risiken für Blutungsereignisse (in %) in der prospektiven GermanVasc-Kohortenstudie anhand des GermanVasc-Risikoscores bei Patienten mit chronischer extremitätengefährdender Ischämie.

Projektziel 1b: **Entwicklung eines Vorhersagemodells zum Auftreten von schweren Blutungskomplikationen**

Volltext im Anhang:

Behrendt CA, Kreutzburg T, Nordanstig J, Twine CP, Marschall U, Kakkos S, Aboyans V, Peters F. The OAC3-PAD Risk Score Predicts Major Bleeding Events one Year after Hospitalisation for Peripheral Artery Disease. Eur J Vasc Endovasc Surg. 2022 Mar;63(3):503-510. doi: 10.1016/j.ejvs.2021.12.019. Epub 2022 Feb 4. PMID: 35125278.

Insgesamt sind 81.930 Patienten mit Behandlung der symptomatischen PAVK (47,2% Frauen, mittleres Alter 72,3 Jahre ± 11,1 Jahre) zwischen 1. Januar 2010 und 31. Dezember 2018 in die Analysen eingegangen. Davon wurden 23,0% offen-chirurgisch revaskularisiert, 55,8% wurden endovaskulär interveniert, 7,1% erhielten eine primäre Amputation der unteren Extremitäten und 14,1% wurden konservativ behandelt.

Nach einem Jahr erlitten 2,2% eine Majorblutung (8,4% atraumatisch intrakraniell, 6,7% traumatisch intrakraniell, 25,6% extrakraniell und 59,4% gastrointestinal). Die mediane Follow-up-Zeit betrug 365 Tage. Der OAC3-PAD-Risikoscore identifizierte acht unabhängige Prädiktoren: Orale Antikoagulation vor dem Indexaufenthalt, hohes Alter über 80 Jahre, chronische extremitätengefährdende Ischämie, Herzinsuffizienz, schwere chronische Nierenerkrankung, vorhergehende Blutungsereignisse, Anämie und Demenz (**Abbildung 6**). Die Basischarakteristika der Trainings- und Validierungskohorte finden sich in Tabelle 3. Die Diskriminierung der Risikogruppen in der Validierungskohorte ergab gute Ergebnisse (c = 0,69, 95%-Konfidenzintervall 0,67-0,71).

In den Sensitivitätsanalysen bestätigte sich das Ergebnis mit nur geringfügig schlechterer Diskriminierung nach drei Jahren (c = 0,67, 95%-Konfidenzintervall 0,66-0,68).

Weitere Details finden sich im Anhang (Behrendt CA, et al. 2022).

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Tabelle 3: Basischarakteristika der Trainings- und Validierungskohorte.

	Training cohort N= 49 160	***Validiation cohort N=32 774***
Mittleres Alter in Jahren +- SD	72.3±11.1	72.3±11.2
Weibliches Geschlecht (%)	23164 (47.1)	15498 (47.3)
Ischämische Ruheschmerzen (PAVK III) (%)	5621 (11.4)	3685 (11.2)
Ulcera oder Nekrosen (PAVK IV) (%)	15720 (32.0)	10633 (32.4)
Offen-chirurgisches Vorgehen (%)	11284 (23.0)	7533 (23.0)
Komorbiditäten (Vorjahr)		
Vorgeschichte einer Majorblutung (%)	8025 (16.3)	5346 (16.3)
Herzinsuffizienz (%)	8677 (17.7)	5713 (17.4)
Hypertension (%)	35535 (72.3)	23687 (72.3)
Leberinsuffizienz (%)	1064 (2.2)	683 (2.1)
Übergewicht (%)	4369 (8.9)	2696 (8.2)
Vorgeschichte eines Schlaganfalls (%)	1614 (3.3)	1160 (3.5)
Psychische und Verhaltensstörungen durch Tabak (%)	6568 (15.3)	4388 (13.4)
Alkoholabhängigkeit (%)	1607 (3.3)	1057 (3.2)
Demenz (%)	2958 (6.0)	1951 (6.0)
Höhergradige Niereninsuffizienz (RIFLE 4-5) (%)	2906 (5.9)	2009 (6.1)
Anämie (%)	5182 (10.5)	3509 (10.7)
Vorgeschichte von Krebs (%)	1561 (3.2)	1067 (3.3)
Medikation (Vorjahr)		
NSAIDs (%)	18678 (38.0)	12628 (38.5)
Thrombozytenaggregationshemmer (%)	16282 (33.1)	10780 (32.9)
Orale Antikoagulation (%)	7626 (15.5)	5068 (15.5)
Lipidsenker (%)	22934 (46.7)	15347 (46.8)

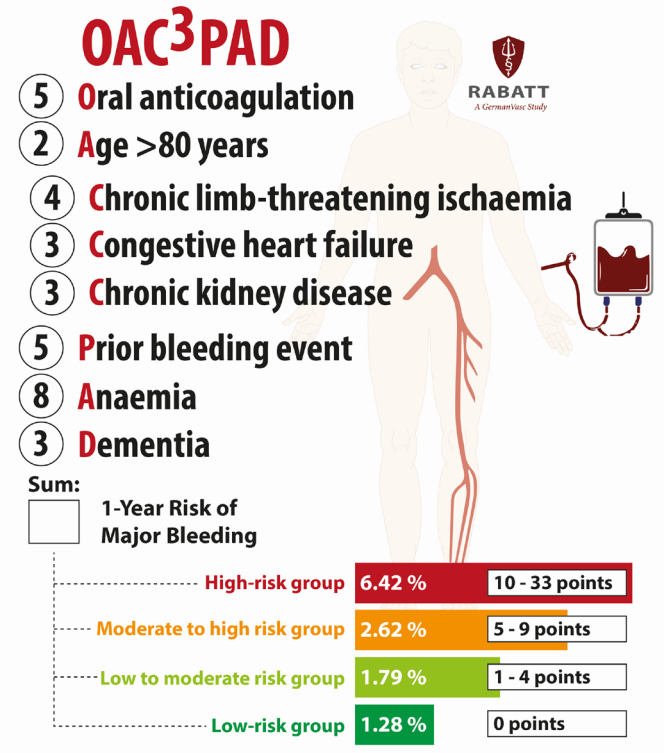

Abbildung 6: Vereinfachende Taschenkarte zum OAC3-PAD-Summenscore zur Vorhersage von schweren Blutungsereignissen bei Patienten mit symptomatischer peripherer arterieller Verschlusskrankheit (PAVK).

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Projektziel 2b: **Externe Validierung des Vorhersagemodells zum Auftreten von schweren Blutungskomplikationen**

Volltext im Anhang:

Peters F, Behrendt CA. External Validation of the OAC3-PAD Risk Score to Predict Major Bleeding Events Using the Prospective GermanVasc Cohort Study. Eur J Vasc Endovasc Surg. 2022 Oct;64(4):429-430. doi: 10.1016/j.ejvs.2022.07.055. Epub 2022 Aug 8. PMID: 35952908.

Zwischen Mai 2018 und Dezember 2021 wurden insgesamt 5.479 Patienten mit symptomatischer PAVK an 37 Gefäßzentren invasiv behandelt, zu denen vollständige Follow-up-Daten vorlagen (33,1% Frauen, mittleres Alter 69,0 Jahre, 67,6% Claudicatio intermittens, 33,0% offen-chirurgische Revaskularisation). Nach einem Jahr erlitten insgesamt 33 Patienten ein Blutungsereignis (1-Jahres-Inzidenz 1,3%). Der OAC3-PAD-Risikoscore teilte insgesamt 46,3% der Patienten der niedrigen Risikogruppe zu (vs. 23,2% niedrig-moderat, 22,6% hoch-moderat, 7,9% hoch). Die beobachteten Blutungsraten waren insgesamt und über alle Gruppen niedriger, als erwartet (v.a. in höheren Risikogruppen). Die Modelldiskrimination war adäquat (c = 0,61, 95%-Konfidenzintervall 0,43-0,80) (**Abbildung 6**). Weitere Details finden sich im Anhang (Peters F, et al. 2022).

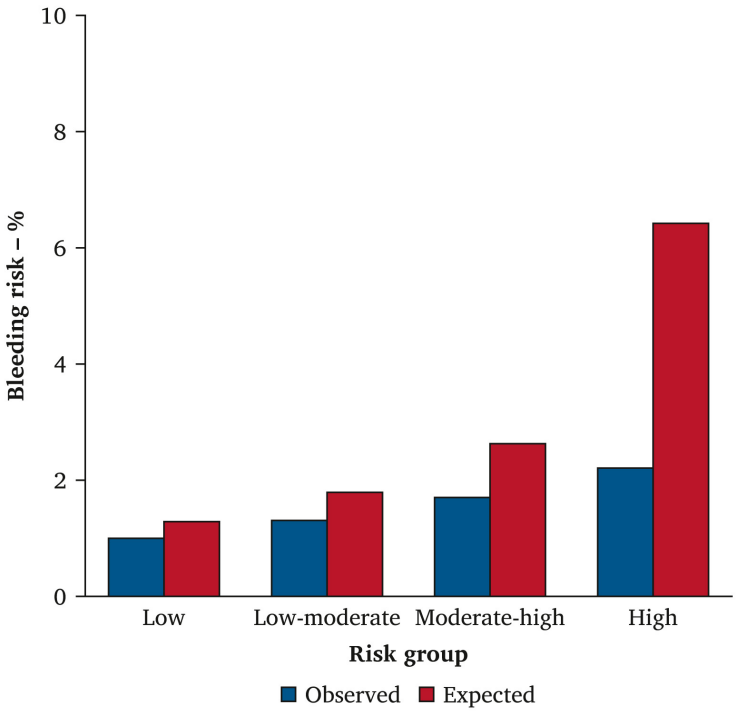

Abbildung 7: Modellkalibrierung zum Vergleich der beobachteten (blau) vs. erwarteten (rot) Risiken für Blutungsereignisse in der prospektiven GermanVasc-Kohortenstudie anhand des OAC3-PAD-Risikoscores.

Projektziel 3: **Befragung zur Nutzung digitaler Gesundheitsanwendungen**

Volltext im Anhang:

Alushi K, Hinterseher I, Peters F, Rother U, Bischoff MS, Mylonas S, Grambow E, Gombert A, Busch A, Gray D, Konstantinou N, Stavroulakis K, Horn M, Görtz H, Uhl C, Federrath H, Trute HH, Kreutzburg T, Behrendt CA. Distribution of Mobile Health Applications amongst Patients with Symptomatic Peripheral Arterial Disease in Germany: A Cross-Sectional Survey Study. J Clin Med. 2022 Jan 19;11(3):498. doi: 10.3390/jcm11030498. PMID: 35159950; PMCID: PMC8836389.

Insgesamt haben 326 Patienten den Fragebogen beantwortet (Antwortrate 96,3%), darunter 34,0% mit Claudicatio intermittens (29,2% Frauen, 70 Jahre im Median) und 66,0% mit

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

chronischer extremitätengefährdender Ischämie (29,5% Frauen, 70 Jahre im Median). Unter allen Teilnehmern hatten 66,8% ein Smartphone und 27,9% benötigten regelmäßige Unterstützung bei der Nutzung des Endgeräts. Insgesamt nutzten 42,5% Smartphone Applikationen und 15% digitale Gesundheitsanwendungen. 19% besaßen Wearables. Etwa ein Fünftel der Befragten antwortete, dass diese Technologien helfen könnten, um den gesunden Lebensstil zu verbessern. In multivariaten Analysen war nur höheres Lebensalter (Odds Ratio 0,89, 95%-Konfidenzintervall 0,86-0,92) mit einer geringeren Wahrscheinlichkeit für den Besitz eines Smartphones assoziiert. Weitere Details finden sich im Anhang (Alushi K, et al. 2022).

Projektziel 3: **Systematische Literaturrecherche zum Einfluss der Ernährung auf Behandlungsergebnisse**

Volltext im Anhang:

Adegbola A, Behrendt CA, Zyriax BC, Windler E, Kreutzburg T. The impact of nutrition on the development and progression of peripheral artery disease: A systematic review. Clin Nutr. 2022 Jan;41(1):49-70. doi: 10.1016/j.clnu.2021.11.005. Epub 2021 Nov 11. PMID: 34864455.

Insgesamt wurden 8502 Artikel identifiziert und gescreent. Nach Evaluation von Titel und Abstract wurden 186 Volltexte analysiert, wonach 82 Studien in die Analysen eingegangen sind (30% randomisierte kontrollierte Studien) (**Abbildung 8**).

Die Nahrungsmittel wurden in Früchte, Gemüse, Antioxidantien, Fette und Öle, Fleisch, Proteine, Vitamine, Ballaststoffe, Spurenelemente, Diäten und Lebensstil eingeteilt.

Die Studien schlossen Kohorten mit einer Größe von acht bis 54.597 Probanden ein, wobei das Alter zwischen 18 und 94 Jahren variierte. Alle eingeschlossenen RCT evaluierten Ernährungssupplemente und nur zwei RCT evaluierten Ernährungsberatungen.

Die Ergebnisse der Studie legten nahe, dass ein Benefit für mediterrane Diät und Nüsse bei Patienten mit PAVK vorliege. Weitere Details finden sich im Anhang (Adegbola A, et al. 2021).

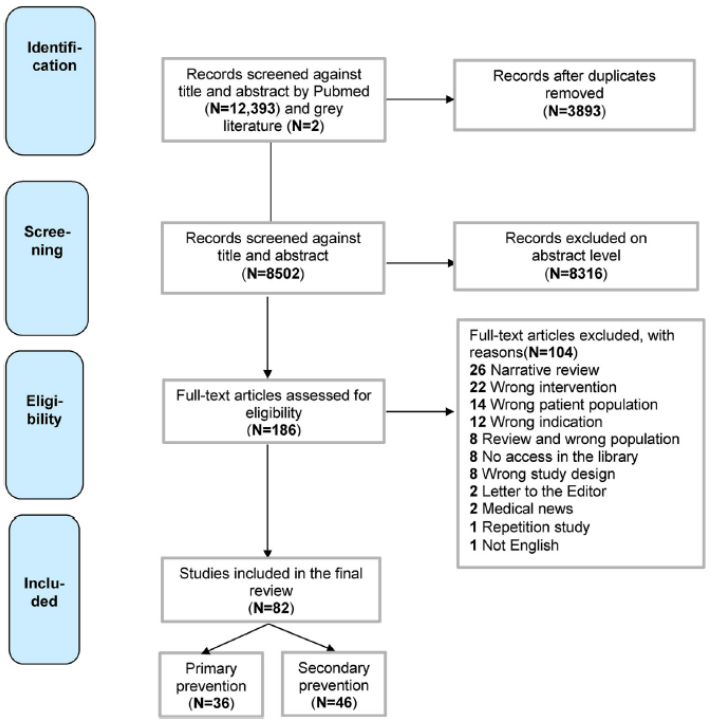

Fig. 1. PRISMA flowchart of the title and abstract and full-text screening about nutritional intake and peripheral arterial occlusive disease.

Abbildung 8: PRISMA-Flowchart der systematischen Literaturrecherche zur Assoziation zwischen Ernährung und Behandlungsergebnis bei Menschen mit peripherer arterieller Verschlusskrankheit (PAVK).

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Projektziel 3 & 4 & 5: **Fokusgruppendiskussionen**

Im Rahmen der moderierten Fokusgruppendiskussionen wurden verschiedene Themenkomplexe diskutiert und konsentiert (per Abstimmung mit absoluter Mehrheitsregel).

a) Als geeignete Endpunkte für die Risikovorhersagemodelle wurden in Routine- und Registerdaten vollständig und valide verfügbare Endpunkte diskutiert. Hierbei wurden vor allem die Erkenntnisse der IDOMENEO-Studie (01VSF16008) ausführlich diskutiert. Die kontextspezifische Validität und Vollständigkeit wurde insbesondere bei Todesfällen/Mortalität und Majoramputationen oberhalb des Knöchels als geeignet identifiziert. Gleichzeitig wurde das Auftreten schwerer Blutungsereignisse als klinisch sehr relevant diskutiert, da die Abwägung bzw. Abstimmung der antithrombotischen Therapie mit Thrombozytenaggregationshemmern, oralen Antikoagulanzien oder deren Kombinationen in der klinischen Praxis sowohl vom Risiko für thromboembolische Ereignisse (z.B. Tod oder Amputation) aber auch Blutungen abhängt.

b) Für die Nutzung der Risikovorhersagemodelle durch Patienten und Angehörige aber auch Ärzte und interessierte Dritte ist die Einrichtung einer webbasierten responsiven Schnittstelle diskutiert worden. Aufgrund zahlreicher rechtlicher und technischer Erwägungen ist eine gerätebasierte Applikation zum Herunterladen und Installieren auf Endgeräten nicht favorisiert worden. Die Teilnehmer der Fokusgruppendiskussion haben allerdings die Möglichkeit zur Nutzung der gerätespezifischen Sensoren und Daten als Vorteil für zukünftige Projekte identifiziert.

c) Die relevanten rechtswissenschaftlichen Fragestellungen werden ausführlich im nächsten Abschnitt beschrieben.

d) Die relevanten rechtswissenschaftlichen Fragestellungen werden ausführlich im nächsten Abschnitt beschrieben.

e) Im Rahmen der Fokusgruppendiskussion zum Thema des privatsphärefreundlichen maschinellen Lernens und möglicher Angriffsszenarien in der klinischen Anwendung konnten verschiedene Aspekte ausführlicher erläutert werden. Den Informationswissenschaftlern wurde hierbei ein detaillierter Einblick in die Nutzung faktisch anonymisierter Forschungsdaten über das Wissenschafts-Data-Warehouse der BARMER gewährt und anschließend die Anwendung derartiger Risikovorhersagemodelle in der klinischen Praxis demonstriert. Ein besonderer Diskussionspunkt war die Umsetzbarkeit derartiger Projekte vor dem Hintergrund der EU-Datenschutzgrundverordnung.

Bei der Fokusgruppendiskussion zum Umfang von evidenzbasierten Empfehlungen zum sogenannten Best Medical Treatment bei Patienten mit PAVK wurden die aktuell in Erstellung oder Überarbeitung befindlichen nationalen und internationalen Praxisleitlinien diskutiert. Durch die Projektleitung wurde dabei Einblick in die systematischen Literaturrecherchen gewährt, die sich mit den Kernthemen beschäftigten. Als Evidenzlücke in bisherigen Leitlinien wurde das Thema Ernährung identifiziert, weshalb eine eigene systematische Literaturrecherche angestoßen wurde. Die gesammelten Empfehlungen wurden in Form eines edukativen laienverständlichen Übersichtsartikels gemeinsam mit einer deutschlandweiten Patientenvertretung zusammengefasst.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Projektziel 3: **Privatsphärefreundliches maschinelles Lernen und Angriffsszenarien in der klinischen Praxis**

Volltext im Anhang:
Stock, J., Petersen, T., Behrendt, CA. et al. Privatsphärefreundliches maschinelles Lernen. Informatik Spektrum 45, 70–79 (2022). https://doi.org/10.1007/s00287-022-01438-3
Stock, J., Petersen, T., Behrendt, CA. et al. Privatsphärefreundliches maschinelles Lernen. Informatik Spektrum 45, 137–145 (2022). https://doi.org/10.1007/s00287-022-01440-9

Der Einsatz von Künstlicher Intelligenz (KI) erfreut sich seit einigen Jahren immer größerer Beliebtheit, nicht zuletzt dank sinkender Kosten für Rechenleistung. In diesem Dokument wird KI als Synonym zum Maschinellen Lernen (ML), einer Klasse selbstadaptiver Algorithmen, verstanden. Das adaptive Prozess der Algorithmen besteht aus einer Lern- bzw. Trainingsphase, die in der Regel dem Einsatz eines KI-Systems vorangeht. Nach einer erfolgreichen Trainingsphase können KI-Modelle dafür verwendet werden, anhand bisher unbekannter Daten Vorhersagen zu treffen. Je nach Einsatzgebiet können diese Vorhersagen unterschiedlicher Natur sein, beispielsweise:

- Verhaltensvorhersagen (Rückzahlungswahrscheinlichkeit von Krediten, Rückfälligkeit von Straftätern, ...),
- Mustererkennung (Objekterkennung auf Fotos, Spracherkennung, Gesichtserkennung, Charaktereigenschaften anhand von Gesichtszügen, ...),
- Medizinische Entscheidungshilfen (Datenanalyse einer Patientenakte, Krebsverdacht anhand von MR-/CT-Scans, ...).

Rechtssicherheit für die Verwendung von KI-Systemen ist nicht immer gegeben. Die Datenschutz-Grundverordnung (DSGVO) gibt zwar allgemeine Regeln für den Einsatz von Algorithmen vor (vor allem bezüglich automatisierter Entscheidungen in Art. 22 DSGVO), allerdings ist die Auslegung einzelner Passagen für KI-Systeme unklar. Das große Potential sinnvoller KI-Anwendungen kann sich erst durch die Schaffung von Rechtssicherheit abrufen lassen.

Die im Projekt entstandenen Veröffentlichungen (**siehe Anhang**) erklären die allgemeinen Grundlagen maschineller Lernverfahren, wie beispielsweise gängige Lerntypen oder auftretende Verzerrungen beim Einsatz von ML-Verfahren. Darin stellen die Konsortialpartner grundlegende Klassen maschineller Lernverfahren vor und geben einen beispielhaften Einblick in häufig verwendete Techniken. Es wird das Prinzip des erklärbaren maschinellen Lernens dargestellt und die Autoren beschreiben verschiedene Angriffe auf ML-Modelle. Übergeordnet stellen die Konsortialpartner schließlich verschiedene Verfahren für privatsphäre-freundliches maschinelles Lernen dar und illustrieren den Einsatz von ML-Anwendungen im Gesundheitswesen. Auf diesen Hintergrundkapiteln aufbauend werden in Thesen zu rechtlichen Aspekten des ML aufgestellt. Diese sollen als interdisziplinäre Diskussionsgrundlage dienen und im Laufe des Diskurses fortwährend ergänzt, konkretisiert und mit Rechtsgrundlagen untermauert werden. Ziel ist die Prüfung der Thesen hinsichtlich ihrer Haltbarkeit und gegebenenfalls übersehener Implikationen. Falls Regelungsdefizite vorhanden sind, sollen diese explizit benannt werden.

Angesichts des großen aktuellen gesellschaftlichen Diskurses und der teils unklaren Rechtsauslegung sind KI-Systeme überwiegend als „neue Technologie“ zu bewerten. Damit wird bereits ein Kriterium für die Notwendigkeit der Anfertigung einer Datenschutzfolgeabschätzung (DSFA) erfüllt. Mit dem oft datenintensiven Training von KI-Algorithmen wird weiterhin der Bestand der „systematischen“ und „umfangreichen“ Verarbeitung von vielen KI-Systemen erfüllt. Werden dabei beispielsweise besonders sensible Daten gemäß Art. 9 Abs. 1 DSGVO verarbeitet, dürfte die Erstellung einer DSFA in vielen Fällen unumgänglich sein.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Eine Schwierigkeit, die sich durch die teils schwerwiegenden Konsequenzen durch den Einsatz von ML-Technologien ergibt, ist, dass die DSGVO großteils individuelle Betroffenenrechte und Pflichten auf Seiten der verarbeitenden Stellen betrachtet. Allerdings sind – je nach Anwendungsbiet – mögliche Auswirkungen auf gesamtgesellschaftliche Rechte und Freiheiten in einer Folgenabschätzung oft mindestens genauso relevant und untersuchenswert. So plädiert beispielsweise auch die Datenethikkommission in ihrem Gutachten je nach Schädigungspotenzial eines KI-Systems für eine Abschätzung der „Risiken für Selbstbestimmung, Privatheit, körperliche Unversehrtheit, persönliche Integrität sowie für Vermögen, Eigentum, und Gleichbehandlung“. Ferner sind zahlreiche zivilgesellschaftliche und öffentliche Organisationen zurzeit damit beschäftigt, geeignete Rahmenwerke und umfassende Verfahren zu entwickeln.

Projektziel 4: **Haftungsrechtliche Ergebnisse**

Volltext im Anhang:
Schmidt J. Die Auswirkungen der Nutzung von KI-Software auf die ärztliche Haftung. GesR. 2023;341-353.

Vor dem Hintergrund neuer technologischer Möglichkeiten stellt sich auch im Bereich des zivilen Medizinrechts die Frage, wie das Recht mit der rasanten Entwicklung Schritt halten kann. Der haftungsrechtliche Projektteil setzt sich vertieft mit der Frage auseinander, welche Auswirkungen die Nutzung von KI-Software auf die ärztliche Haftung hat. Hierbei werden neben dem rechtlichen Status Quo auch die aktuellen Reformvorhaben in den Blick genommen. Dem Text liegt ein Gutachten im Rahmen des vom Gemeinsamen Bundesausschuss finanzierten Forschungsprojekts RABATT - Risikoscore für eine Algorithmenbasierte Behandlerunabhängige Aufklärung zum Therapieerfolg und zur Therapieempfehlung zugrunde.

Die Nutzung von algorithmenbasierten Anwendungen wirft einige rechtliche Fragen auf, von denen teilweise unklar ist, ob sie durch die bestehenden Normen zufriedenstellend beantwortet werden können. Dies gilt insbesondere für das Haftungsrecht. Bei Schädigungen eines Patienten im Rahmen einer medizinischen Behandlung stellt sich häufig die Frage, ob der behandelnde Arzt nicht einen Fehler bei der Therapie gemacht hat, die den Schaden verursacht hat und für den er zu haften hat. Diese Haftung stellt die Kehrseite der grundsätzlich bestehenden Therapiefreiheit des Arztes dar und begrenzt diese. Unklar ist allerdings, welche Folgen es für die ärztliche Haftung hat, wenn nicht dieser selbst die Behandlungsentscheidung trifft, sondern diese maßgeblich durch eine KI-basierte Software getroffen wird, die durch Auswertung einer großen Zahl von Daten gezielt auf den individuellen Patienten zugeschnittene Therapievorschläge macht. Aufgrund der hohen Anzahl der ausgewerteten Daten kann ein fehlerfrei programmierter Algorithmus eine höhere Entscheidungsevidenz für sich beanspruchen, als ein einzelner Mediziner, dessen Entscheidungen auf erworbener Erfahrung und Wissen aus Lektüre von Fachliteratur fußen. Allerdings ist es für den einzelnen Nutzer in aller Regel nicht erkennbar, wie ein Algorithmus zu seinem Ergebnis gekommen ist (Blackbox-Effekt). Dies gilt insbesondere für KI-basierte, selbstlernende Algorithmen. Es stellt sich insofern die Frage, wie es um die Therapiefreiheit eines Arztes bestellt ist, dem man einen solchen Algorithmus an die Seite stellt. Zudem ist zu klären, ob das derzeitige Haftungsregime des Arzthaftungsrechts die Verantwortungsverteilung in einer solchen Behandlungssituation noch korrekt abbildet und welche Haftungsfolgen sich aus dem Einsatz einer algorithmenbasierten Software in der täglichen Arbeit ergeben.

Der Aufsatz im Anhang befasst sich mit den arzthaftungsrechtlichen Aspekten bei der Verwendung von algorithmenbasierten Diagnose- und Therapietools. Die ärztliche Haftung für Pflichtverletzungen gegenüber Patienten kann sich aus Vertrags- und Deliktsrecht ergeben. Im

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Aufsatz werden zunächst die Grundzüge der ärztlichen Haftung dargestellt, um im Weiteren auf die Besonderheiten der ärztlichen Haftung bei Nutzung von selbstlernenden Algorithmen bei der Behandlung einzugehen. Hierfür sind zunächst die Begrifflichkeiten KI und Algorithmus, Blackbox-Effekt und Intelligenzrisiko zu klären. Im Anschluss wird die grundsätzliche ärztliche Haftung für Patientenschäden kurz skizziert. Darauf aufbauend wird analysiert, welche Folgen die Verwendung von algorithmenbasierten Diagnose- und Therapietools nach dem geltenden Recht hat. Hier-bei wird ein besonderer Fokus auf die Frage gelegt, welche Auswirkungen algorithmenbasierte Tools auf den „ärztlichen Standard“ haben können. Weiterhin wird auch die Frage analysiert, ob die Verwendung von algorithmenbasierten Tools Einfluss auf die Anwendung der Beweislastregeln des Arzthaftungsrechts haben kann. Zudem wird auch der Einfluss der Nutzung solcher Tools auf die ärztliche Aufklärungspflicht nach § 630e BGB thematisiert. Zuletzt wird ein Ausblick auf die mögliche zukünftige Rechtsentwicklung gegeben.

Die rechtswissenschaftliche Beurteilung der haftungsrechtlichen Fragestellungen hat nachfolgende Ergebnisse erbracht:

- Die Nutzung von KI-Software im medizinischen Bereich stellt nicht per se eine Pflichtverletzung dar.
- KI-Software hat das Potential, in Zukunft den ärztlichen Standard zu prägen. Zurzeit ist die Nutzung von KI-Software zur Behandlung und Diagnostik jedoch als Neulandmethode einzustufen.
- Sollte KI-Software in Zukunft den ärztlichen Standard prägen, so würde sich aller Voraussicht nach auch die Standardbestimmung, weg von Empfehlungen von Behandlungsmethoden von Krankheitsbildern, hin zu Handlungsempfehlungen bezogen auf bestimmte Patientengruppen, entwickeln. Eine Nichtnutzung von KI-Software, die zum Standard geworden ist, dürfte erst nach einer Übergangsphase geeignet sein, um zu einer Pflichtverletzung zu führen.

Projektziel 5: **Sozialrechtliche Ergebnisse**

Dissertationstext in Erstellung:
Schneller R. Normgesteuerte Rezeption medizinischer Erkenntnisse im SGB V.

Im sozialrechtlichen Bereich stellt sich die Frage, wie das Recht mit modernen Formen der medizinischen Wissensgenerierung umgehen kann. Dieser Projektteil setzt sich vertieft mit der Frage auseinander, welche Anforderungen an medizinisches Wissen sich aus unterschiedlichen sozialrechtlichen Normen ableiten lassen und wann die gerichtliche Praxis diese Anforderungen erfüllt sieht.

Die übergreifende Frage, ob eine Ermittlung des medizinischen Erkenntnisstandes unter Verwendung von Big Data gestützten Technologien dem Gesetzeswortlaut entspricht, hängt dabei zentral von der Offenheit der herangezogenen Normen für unterschiedliche technologische Ansätze ab. Weil die einschlägigen Normen im SGB V teils seit Inkrafttreten des Gesetzes im Jahr 1989 in diesen Punkten unverändert bestehen, geht die Dissertation chronologisch vor. So konnte herausgearbeitet werden, dass der Gesetzgeber zwar ursprünglich eine konsensbasierte Bewertung medizinischen Wissens im Sinn hatte, zwischenzeitlich von der Rechtswissenschaft aber auch Bezüge auf ärztliche Leitlinien und die etablierte evidenzbasierte Medizin für die Auslegung der Norm herangezogen werden. Trotzdem wird auch deutlich, dass die Grenzen dessen, was der Gesetzgeber mit den untersuchten Normen bezwecken wollte und gemeint haben konnte, zunehmend erreicht und teilweise bereits jetzt überspannt werden. Die Normen sind damit zwar grundsätzlich technologieoffen, was insbesondere auch eine Wissensgenerierung unter Verwendung von Big Data einbeziehen könnte, die Ambiguität des Gesetzes und die Zurückhaltung des Gesetzgebers fördert aber keine einheitliche Handhabe des Rechts mit diesen Technologien,

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

wodurch unter dem derzeitigen Rechtsrahmen weiterhin Unsicherheiten über die konkreten Anforderungen an die medizinische Wissensqualität bestehen.

Zur Auflösung der gesetzlichen Ambiguität befasst sich die Dissertation mit der medizinischen Wissensbewertung in der Praxis am Beispiel einzelner Methodenbewertungsverfahren des G-BA und des IQWiG. Unter Verwendung der Ergebnisse der Fokusgruppendiskussionen (oben Projektziel 3, 4 und 5) konnte so die Komplexität moderner medizinischer Wissensgenerierung und -Bewertung für den rechtswissenschaftlichen Diskurs sichtbar gemacht werden. Hierdurch hat sich gezeigt, dass die in der deutschen Rechtswissenschaft üblichen Verweise auf den Gesetzestext oder die einschlägigen Verfahrensordnung zur qualitativen Beschreibung eines medizinischen Erkenntnisstandes regelmäßig zu unterkomplex sind, um die tatsächliche medizinwissenschaftliche Praxis sachgerecht zu beschreiben. Der Blick in die medizinische Praxis wirft aber auch Fragen zur Handhabe der rechtlichen Praxis auf: wie ermittelt und bewertet diese einen medizinischen Erkenntnisstand und für welche neueren Entwicklungen der Medizin ist sie hierbei offen?

Zur Beantwortung dieser Folgefragen geht die Dissertation auf Leitentscheidungen des BSG ein. Übergreifende Frage ist auch hier, welche qualitativen Anforderungen die Sozialgerichte an das medizinische Wissen stellen, um herausarbeiten zu können, mit welchen Technologien die Rechtsprechung wie umgeht. Historisch lässt sich gut nachvollziehen, wie die Rechtsprechung von der früher konsensbasierten Praxis zwischenzeitlich zu Verweisen auf die ärztlichen Leitlinien übergangen ist und heutzutage bei den Verfahren der evidenzbasierten Medizin angekommen ist. Moderne Techniken der Wissensgenerierung wie Big Data, künstliche Intelligenz oder Formen des maschinellen Lernens wurden in der sozialrechtlichen Rechtsprechung zur Bewertung eines medizinischen Erkenntnisstandes bisher jedoch noch nicht thematisiert. Vielmehr zeigt sich, dass schon die Verwendung medizinischer Literatur ohne diese technischen Entwicklungen die Rechtspraxis vor methodische Herausforderungen stellt.

Die Rechtsprechung zeigt damit zwar im Umgang mit den Normen eine grundsätzliche Innovations- und Technologieoffenheit. Der bisherige Umgang der Gerichte mit medizinischen Quellen wirft aber Zweifel auf, ob die Rechtsprechung Erkenntnisse aus neueren Technologien einordnen kann. Erschwerend dürfte hier hinzutreten, dass die Sozialgerichte mit den technologischen Entwicklungen der Medizin bisher wenige Berührungspunkte haben; die Fragen zur Richtigkeit der Aussagen von KI-Modellen wurden bislang vorwiegend im Datenschutzrecht thematisiert und werden erst jüngst im Kontext der Ermittlung des sozialrechtlichen „Stand der medizinischen Erkenntnis“ diskutiert. Diese Verwendung von künstlicher Intelligenz als Instrument der Wissensgenerierung ist in der sozialgerichtlichen Rechtsprechung bisher jedoch noch nicht behandelt worden.

Der letzte Teil der Untersuchung befasst sich mit einem rechtsmethodischen Rahmen zur Bewertung medizinischer Erkenntnisse. Weil die Gesetzgebung und das Vorgehen der Rechtspraxis gezeigt haben, dass schon der Umgang mit konventionellen medizinischen Quellen, also solchen ohne Bezug zu Big Data und ähnlichen Technologien, erhebliche methodische Probleme für die Rechtswissenschaft bereitet, fokussiert sich die in der Untersuchung entwickelte Methodik zunächst auf diese. Hierbei werden verschiedene etablierte medizinische Bewertungsinstrumente wie AGREE, CONSORT und PRISMA darauf untersucht, ob einzelne Bewertungskriterien dieser Instrumente auch ohne vertiefte medizinische Kenntnisse anwendbar sind. Hieraus wird eine rechtswissenschaftliche Methodik zur Handhabung medizinischer Literatur entwickelt. Diese Methodik ermöglicht eine einheitliche und sachgerechte Einbindung medizinischen Wissens in rechtliche Entscheidungen. Trotzdem werden Grenzen deutlich, die weiterer Untersuchungen bedürfen: die Bedeutung KI-gestützter Wissensgenerierung in der Medizin dürfte die nächsten Jahre nur weiter zunehmen. In diesem Punkt gilt es zu beobachten, wie die Rechtsprechung mit diesen Technologien umgehen wird, um die in der Dissertation entwickelte Methodik weiter anzupassen. Auch wird deutlich, dass der Gesetzgeber die Normen, die auf einen

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

medizinischen Erkenntnisstand verweisen, dringend überarbeiten sollte, um eine anhaltende Technologieoffenheit des Rechts sicherzustellen. Hierfür werden in der Untersuchung verschiedene Vorschläge diskutiert, die aber davon abhängen, ob der Gesetzgeber selbst tätig wird, das Bundesgesundheitsministerium zu Verordnungen ermächtigt, oder die Frage der Handhabungen der Technologien für die sozialrechtlichen Vorgaben an den G-BA delegiert.

Die sozialrechtliche Untersuchung hat damit folgende Ergebnisse erbracht:

- Das SGB V verweist an vielen Stellen auf einen medizinischen Erkenntnisstand. Diese Normen sind grundsätzlich offen für neues medizinisches Wissen.
- Angesichts des Alters der Normen ist zu bezweifeln, dass der Gesetzgeber beliebig innovative medizinische Wissensgenerierungs- und Wissensbewertungsverfahren gemeint haben kann.
- Die Rechtskonformität der neuen Technologien hängt zentral davon ab, wie innovativ diese sind; desto innovativer sie sind, desto eher ist eine Anpassung des Rechtsrahmens erforderlich.
- Moderne Techniken der Wissensgenerierung wie Big Data, künstliche Intelligenz oder Formen des maschinellen Lernens wurden in der sozialrechtlichen Rechtsprechung zur Bewertung eines medizinischen Erkenntnisstandes nicht thematisiert.
- Die Rechtsanwendung steht bereits vor Herausforderungen im Umgang mit konventioneller medizinischer Literatur. Diese Herausforderungen lassen sich durch eine Übersetzung medizinischer Bewertungsinstrumente in die Rechtsmethodik lösen.

Die sozialrechtliche Untersuchung befasst sich damit nicht im Schwerpunkt mit dem Datenschutzrecht. Hierzu lässt sich aber auf die Dissertation von Lea Köttering mit dem Titel „Datenschutzrechtliche Regulierung maschineller Lernverfahren, 2024“ verweisen, die sich derzeit noch im Verfahren der Begutachtung befindet. Diese Untersuchung beschäftigt sich ausführlich und durchaus grundlegend mit dem Verhältnis von maschinellen Lernverfahren und den datenschutzrechtlichen Regulierungen (DSGVO und sonstige Gesundheitsdatenschutzregelungen). Ausgangspunkt ist dabei die Frage, ob, wie oft angenommen, die datenschutzrechtliche Regulierung die Entwicklung und Anwendung eines maschinellen Lernverfahrens im Gesundheitswesen blockiert oder doch zumindest erschwert. Die Arbeit unterzieht den regulatorischen Rahmen einer sehr differenzierten Analyse, die praktische Fragen ebenso wie die Einbettung der Systeme einbezieht und die zugleich Grenzen eines datenschutzrechtlichen Ansatzes bestimmt. Dabei wird auch auf den Entwurf des Gesundheitsdatennutzungsgesetzes eingegangen, soweit dieser für die Beurteilung der anstehenden Fragen von Bedeutung ist.

7. Diskussion der Projektergebnisse

Das RABATT-Konsortium hat vom 1. April 2019 bis zum 31. März 2023 verschiedene datenbasierte und qualitative Forschungsprojekte durchgeführt und dabei sowohl gefäßmedizinische Versorgungsforscher als auch Informations- sowie Rechtswissenschaftler vernetzt, um verschiedene Schnittstellenfragen an Anwendungsbeispielen zu bearbeiten.

Zu den wesentlichen Projektergebnissen zählten dabei die Entwicklung und Validierung von zwei Risikovorhersagemodellen unter Einbeziehung großer Datenquellen aus Routinedaten und Registern sowie maschineller Lernverfahren.[13, 14] Der GermanVasc-Risikoscore kann demnach genutzt werden, um das Risiko für Majoramputation und Tod innerhalb von fünf Jahren in der Zielpopulation vorherzusagen. Der OAC3-PAD-Risikoscore ist dagegen geeignet, um das Risiko für schwere Blutungskomplikationen innerhalb eines Jahres vorherzusagen. In der Kombination beider Vorhersagemodelle kann damit in der klinischen Alltagspraxis eine patientenindividuelle Abstimmung der gerinnungshemmenden antithrombotischen Therapie unterstützt werden. Diese alltägliche Herausforderung ist rein zahlenmäßig bereits relevant;

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

etwa 24% der Menschen zwischen 45 und 74 Jahren in Deutschland haben per Definition eine PAVK, wobei ca. 1 Mio. gesetzlich Versicherte Patienten in Krankenhäusern eine entsprechende Haupt- oder Nebendiagnose aufweisen.[26, 27]

Für die Entwicklung der Vorhersagemodelle wurden routinemäßig erhobene faktisch anonymisierte Forschungsdaten der BARMER verwendet, die zuvor primär für administrative Zwecke erhoben wurden und zunehmend für die Beantwortung wissenschaftlicher Fragestellungen verwendet werden.[28] Durch die im RABATT-Projekt durchgeführten Projekte und klinischen Anwendungsbeispiele ergeben sich verschiedene Fragestellungen an den Schnittstellen zur Rechtswissenschaft und Informatik.

Aus methodischer Sicht ist die Frage der internen und externen Validität ganz grundsätzlich zu betrachten, nicht nur bei Beobachtungsstudien mit Register- oder Routinedaten. Jede systematische bzw. strukturierte Sammlung von Daten in Registern, ob primär zu wissenschaftlichen oder administrativen Zwecken, erfordert eine kontextspezifische Validierung, um deren Güte zu beurteilen.[28, 29] Die Qualität von Trainingsdaten, anhand deren Vorhersagemodelle mithilfe von maschinellen Lernverfahren bzw. künstlicher Intelligenz entwickelt werden, ist in letzter Zeit zunehmend in den Fokus gerückt, weil sich aus verzerrten bzw. falschen Vorhersagemodellen ganz grundsätzliche Implikationen für rechtliche Bewertungen ergeben.[30] So kann die falsche Vorhersage von Risiken potenziell zu falschen Behandlungen und Schäden für die Betroffenen führen, wodurch sich Fragen aus dem Bereich des Haftungsrechts ergeben. Gleichzeitig ergeben sich technische Möglichkeiten der Einflussnahme auf Vorhersagemodelle und Anwendungen der künstlichen Intelligenz, was für Angriffe durch Dritte missbräuchlich verwendet werden kann. Dagegen erscheint der Aspekt des Datenschutzes und des privatsphärefreundlichen maschinellen Lernens fast schon trivial.[21, 22] Dies gilt insofern auch für die Nutzerbefragung im RABATT-Projekt, bei dem weniger als 20 unter bis zu 700 Einrichtungen mit gefäßmedizinischen Abteilungen, Kliniken oder Sektionen teilgenommen haben. Die Repräsentativität der erhobenen Daten muss daher in zukünftigen Studien bestätigt werden.

Im RABATT-Projekt konnte eine prospektiv für das IDOMENEO-Projekt erhobene und extern unabhängig qualitätsgesicherte Datenbasis mit bis zu 5.608 Patienten im GermanVasc-Register genutzt werden, um die entwickelten Vorhersagemodelle zu validieren.[6, 7] Die Validierungen ergaben eine moderate Diskrimination zwischen den Risikogruppen, bestätigten allerdings auch, was frühere Validierungsprojekte regelmäßig kritisiert haben: Die heterogene Definition der genutzten Variablen bzw. Risikofaktoren und Endpunkte und die voneinander diskret abweichenden Zielpopulationen führen selten zu einer hervorragenden Modellgüte mit C-Indices über 0,8. Es sei an dieser Stelle aber auch betont, dass die beiden einzigen anderen Vorhersagemodelle zu Blutungsrisiken (HASBLED und REACH) zu deutlich schlechteren Diskriminierungen führten, obwohl diese für kardiologische Indikationen weltweit genutzt werden.[15] Am sinnvollsten wäre aus methodischer Sicht die prospektive Erhebung einer Validierungskohorte mit exakt auf die Trainingskohorte abgestimmten Parametern in ausreichender Fallzahl- und Ereignisgröße. Dieser Aufwand war allerdings im Projekt nicht vorgesehen und erscheint unverhältnismäßig hoch, wenn man darüber hinaus die große Heterogenität der Versorgung in Deutschland mit ca. 670 Kliniken in der PAVK-Behandlung betrachtet.[31]

Die entwickelten Vorhersagemodelle sollten grundsätzlich in der klinischen Praxis auf Anwendbarkeit getestet werden und einer kontinuierlichen Qualitätssicherung und Weiterentwicklung unterzogen werden. Dies kann nur erfolgen, wenn die interdisziplinäre gefäßmedizinische Versorgungsrealität die beiden Risikoscores berücksichtigt.

In den aktuellen Praxisleitlinien der European Society for Vascular Surgery (ESVS) zur antithrombotischen Therapie und zum Management von Menschen mit asymptomatischer PAVK und Claudicatio intermittens wurden die beiden Risikovorhersagemodelle und deren externe Validierungen zur Anwendung bei der Behandlung der Zielpopulation diskutiert bzw.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

als Vorschlag in die Leitlinien integriert.[2, 3] Die internationale Diskussion in Journalen und auf Kongressen hat außerdem zu verschiedenen Validierungsstudien in Frankreich,[16] Niederlande, England, Schweden und den USA geführt; bisher konnte die moderate bis gute Diskrimination des OAC3-PAD und der Subgruppe der chronischen extremitätengefährdenden Ischämie im GermanVasc-Score bestätigt werden. Die online bereitgestellten Kalkulatoren (score.germanvasc.de) sind zudem in verschiedenen Veröffentlichungen und auch in einem Themenheft des Focus Magazins der breiteren Öffentlichkeit präsentiert worden. Ob die konsequente Anwendung dieser Werkzeuge in der patientenindividuellen Behandlung allerdings zu einer Verbesserung des Behandlungsergebnisses beitragen kann, muss in zukünftigen Studien evaluiert werden.

Die im RABATT-Projekt entwickelten und validierten Risikovorhersagemodelle waren allerdings aus der Perspektive der Rechtswissenschaft und Informatik vor allem ein Anwendungsfall zur Bearbeitung komplexer Fragestellungen aus den entsprechenden Fachbereichen. Die konsequente Begleitung und der regelmäßige Austausch zwischen Beteiligten der gefäßmedizinischen Therapie und Versorgungsforschung, der Informatik und Rechtswissenschaft sollte Barrieren überwinden und einen Wissensaustausch fördern. Die Einblicke in die unterschiedlichen Perspektiven und Bewertungsmodelle der drei Bereiche konnten unmittelbar nach dem Start des Projektes zur Identifizierung vieler Schnittstellen führen. Allerdings hat die im Jahr 2020 einsetzende globale Pandemie mit teilweise massiven Restriktionen und Kontaktbeschränkungen, insbesondere im Krankenhaussektor, zu zahlreichen Verzögerungen und Limitationen geführt.

So war die Vor-Ort-Begleitung über weite Teile nicht möglich und auch nicht ethisch vertretbar. Die meisten der Fokusgruppendiskussionen und Workshops mussten über Online-Videokonferenzen durchgeführt werden. Die über längere Episoden primär im Homeoffice durchgeführten Projektanteile haben daher insgesamt einen weniger intensiven Austausch zwischen den Wissenschaftlern erfahren. Da die Begleitung und Bearbeitung von Fragestellungen aus dem rechtlichen Bereich ohnehin als qualitative Forschung nur bedingt objektivierbar war und die Bewertung nicht grundsätzlich auf etablierten Messinstrumenten beruhte, sind die Projektergebnisse in diesem Bereich in weiten Teilen subjektiv geprägt und unterliegen einem Wandel durch zukünftige rechtliche Bewertungen in einer dynamischen Ära der künstlichen Intelligenz.

8. Verwendung der Ergebnisse nach Ende der Förderung

Die Ergebnisse des RABATT-Projekts sollen in der Zukunft kontinuierlich weiterentwickelt werden. Bereits zum Abschluss der Förderdauer sind mehrere Validierungsstudien durch internationale Partner und Forschungsverbünde angestoßen worden, die zum aktuellen Zeitpunkt bereits Vereinfachungen vorgeschlagen haben. Die möglichen Änderungen können zukünftig auch in den webbasierten Kalkulator eingefügt werden, der mit dem Ende der Förderdauer durch das Deutsche Institut für Gefäßmedizinische Gesundheitsforschung gGmbH in Berlin weiterbetrieben wird. Gleichzeitig hat sich das Forschungsinstitut der Deutschen Gesellschaft für Gefäßchirurgie und Gefäßmedizin, Gesellschaft für operative, endovaskuläre und präventive Gefäßmedizin e.V. auch bereiterklärt, die Datenbank für Angebote der Prävention, z.B. zu Gefäßsportgruppen oder Rauchentwöhnungsangeboten, zu betreiben. Zum Zeitpunkt der Berichtserstellung ist die lauffähige Version unter „pavk.info“ installiert worden und wird im Jahr 2024 in gemeinsamen Aktionen umworben. Die weitere Bereitstellung der Website und Applikation erfolgt zukünftig durch die vorgenannte Fachgesellschaft und ihr Forschungsinstitut.

Dass die im RABATT-Projekt entstandene laienverständliche Veröffentlichung zu präventivmedizinischen Aspekten in der PAVK-Behandlung auf sehr großes Interesse und Bedarf gestoßen ist (mehr als 5.300 Downloads in 2 Jahren), ist ermutigend. Verschiedene gemeinsame Aktivitäten mit Patientenvertretern und internationalen Forschungsverbünden

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

sind bereits entstanden, um evidenzbasierte Informations- und Aufklärungstexte in verschiedenen Sprachen für die breite Öffentlichkeit zu entwickeln. Während sich allerdings viele dieser Aktivitäten primär an Fachleute aus gefäßmedizinischen Fachdisziplinen richten (z.B. Gefäßchirurgie, interventionelle Radiologie, Angiologie), besteht eine Herausforderung darin, die Betroffenen und deren primärbehandelnde Ärzte im ambulanten Sektor zu erreichen. Hierfür wäre etwa die Nutzung des Risikoscores zur Selektion von Hochrisikopopulationen und deren gezielte Kontaktierung mit Informationsbroschüren durch die Krankenkassen denkbar.

9. Erfolgte bzw. geplante Veröffentlichungen

Schwaneberg T, Debus ES, Repgen T, Trute H, Müller T, Federrath H, Marschall U, Behrendt CA. [Development of a learning algorithm using real world evidence data - The RABATT study]. Gefässchirurgie. 2019;24:234-238.

Kreutzburg T, Peters F, Kuchenbecker J, Marschall U, Lee R, Kriston L, Debus ES, Behrendt CA. Editor's Choice - The GermanVasc Score: A Pragmatic Risk Score Predicts Five Year Amputation Free Survival in Patients with Peripheral Arterial Occlusive Disease. Eur J Vasc Endovasc Surg. 2021 Feb;61(2):248-256. doi: 10.1016/j.ejvs.2020.11.013. Epub 2020 Dec 15. PMID: 33334671.

Behrendt CA, Kreutzburg T, Nordanstig J, Twine CP, Marschall U, Kakkos S, Aboyans V, Peters F. The OAC3-PAD Risk Score Predicts Major Bleeding Events one Year after Hospitalisation for Peripheral Artery Disease. Eur J Vasc Endovasc Surg. 2022 Mar;63(3):503-510. doi: 10.1016/j.ejvs.2021.12.019. Epub 2022 Feb 4. PMID: 35125278.

Peters F, Behrendt CA. External Validation of the OAC3-PAD Risk Score to Predict Major Bleeding Events Using the Prospective GermanVasc Cohort Study. Eur J Vasc Endovasc Surg. 2022 Oct;64(4):429-430. doi: 10.1016/j.ejvs.2022.07.055. Epub 2022 Aug 8. PMID: 35952908.

Behrendt CA, Rother U, Uhl C, Goertz H, Stavroulakis K, Gombert A; die Kommission PAVK und DFS der DGG e. V.. Vorhersage von schweren Blutungsereignissen bei Patienten mit peripherer arterieller Verschlusskrankheit: Der OAC3-PAD-Risikoscore [Predicting major bleeding events in patients with peripheral arterial disease: the OAC3-PAD risk score]. Gefasschirurgie. 2022;27(3):208-212. German. doi: 10.1007/s00772-022-00881-6. Epub 2022 Mar 11. PMID: 35291723; PMCID: PMC8913852.

Rosenberg, Y., Görtz, H., Rother, U. et al. Empfehlungen zur konservativen Therapie und Sekundärprävention der peripheren arteriellen Verschlusskrankheit (PAVK): Eine evidenzbasierte Informationsbroschüre für Betroffene. Gefässchirurgie 27, 39–45 (2022). https://doi.org/10.1007/s00772-021-00855-0.

Adegbola A, Behrendt CA, Zyriax BC, Windler E, Kreutzburg T. The impact of nutrition on the development and progression of peripheral artery disease: A systematic review. Clin Nutr. 2022 Jan;41(1):49-70. doi: 10.1016/j.clnu.2021.11.005. Epub 2021 Nov 11. PMID: 34864455.

Peters F, Behrendt CA; IDOMENEO Collaborators. Limb Related Outcomes of Endovascular vs. Open Surgical Revascularisation in Patients with Peripheral Arterial Occlusive Disease: A Report from the Prospective GermanVasc Cohort Study. Eur J Vasc Endovasc Surg. 2023 Jul;66(1):85-93. doi: 10.1016/j.ejvs.2023.03.040. Epub 2023 Mar 25. PMID: 36972814.

Stock, J., Petersen, T., Behrendt, CA. et al. Privatsphärefreundliches maschinelles Lernen. Informatik Spektrum 45, 70–79 (2022). https://doi.org/10.1007/s00287-022-01438-3

Stock, J., Petersen, T., Behrendt, CA. et al. Privatsphärefreundliches maschinelles Lernen. Informatik Spektrum 45, 137–145 (2022). https://doi.org/10.1007/s00287-022-01440-9

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Alushi K, Hinterseher I, Peters F, Rother U, Bischoff MS, Mylonas S, Grambow E, Gombert A, Busch A, Gray D, Konstantinou N, Stavroulakis K, Horn M, Görtz H, Uhl C, Federrath H, Trute HH, Kreutzburg T, Behrendt CA. Distribution of Mobile Health Applications amongst Patients with Symptomatic Peripheral Arterial Disease in Germany: A Cross-Sectional Survey Study. J Clin Med. 2022 Jan 19;11(3):498. doi: 10.3390/jcm11030498. PMID: 35159950; PMCID: PMC8836389.

Behrendt CA, Sedrakyan A, Katsanos K, Nordanstig J, Kuchenbecker J, Kreutzburg T, Secemsky EA, Debus ES, Marschall U, Peters F. Sex Disparities in Long-Term Mortality after Paclitaxel Exposure in Patients with Peripheral Artery Disease: A Nationwide Claims-Based Cohort Study. J Clin Med. 2021 Jul 2;10(13):2978. doi: 10.3390/jcm10132978. PMID: 34279461; PMCID: PMC8268810.

Kotov A, Blasche DA, Peters F, Pospiech P, Rother U, Stavroulakis K, Remig J, Schmidt-Lauber C, Zeller T, Görtz H, Teßarek J, Behrendt CA. The Impact of Chronic Kidney Disease on Mid-Term Outcomes after Revascularisation of Peripheral Arterial Occlusive Disease: Results from a Prospective Cohort Study. J Clin Med. 2022 Aug 14;11(16):4750. doi: 10.3390/jcm11164750. PMID: 36012989; PMCID: PMC9409847.

Rosenberg Y, Behrendt CA. Best medical treatment in patients with PAD. Vasa. 2023 Sep;52(5):293-301. doi: 10.1024/0301-1526/a001076. Epub 2023 May 8. PMID: 37151024.

Peters F, Kreutzburg T, Kuchenbecker J, Marschall U, Remmel M, Dankhoff M, Trute HH, Repgen T, Debus ES, Behrendt CA. Behandlungsqualität in der operativ-interventionellen Gefäßmedizin – was können Routinedaten der Krankenkassen leisten? Gefässchirurgie. 2020;25:530-540. DOI: 10.1007/s00772-020-00664-x.

Peters F, Kreutzburg T, Kuchenbecker J, Marschall U, Remmel M, Dankhoff M, Trute HH, Repgen T, Debus ES, Behrendt CA. [Quality of care in surgical interventional vascular medicine: what can be achieved with routinely collected data from health insurance providers?]. Gefässchirurgie. 2020;25:19-28. DOI: 10.1007/s00772-020-00679-4.

Peters F, Kreutzburg T, Kuchenbecker J, Debus ES, Marschall U, L'Hoest H, Behrendt CA. A retrospective cohort study on the provision and outcomes of pharmacological therapy after revascularization for peripheral arterial occlusive disease: A study protocol. BMJ Surgery, Interventions & Health Technologies 2020;2:e000020. doi: 10.1136/bmjsit-2019-000020.

Kreutzburg T, Peters F, Rieß HC, Hischke S, Marschall U, Kriston L, L´Hoest H, Sedrakyan A, Debus ES, Behrendt CA. Editor´s Choice - Comorbidity Patterns among Patients with Peripheral Arterial Occlusive Disease in Germany – A Trend Analysis of Health Insurance Claims Data. Eur J Vasc Endovasc Surg. 2020;59:59-66.

Schmidt J. Die Auswirkungen der Nutzung von KI-Software auf die ärztliche Haftung. GesR. 2023;341-353.

10. Literaturverzeichnis

1. Gerhard-Herman MD, Gornik HL, Barrett C, Barshes NR, Corriere MA, Drachman DE*, et al.* 2016 AHA/ACC Guideline on the Management of Patients With Lower Extremity Peripheral Artery Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2017;**69**:e71-e126.

2. Twine CP, Kakkos SK, Aboyans V, Baumgartner I, Behrendt CA, Bellmunt-Montoya S*, et al.* Editor's Choice - European Society for Vascular Surgery (ESVS) 2023 Clinical Practice Guidelines on Antithrombotic Therapy for Vascular Diseases. *Eur J Vasc Endovasc Surg* 2023;**65**:627-89.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

3. Nordanstig J, Behrendt CA, Baumgartner I, Belch J, Bäck M, Fitridge R*, et al.* European Society for Vascular Surgery (ESVS) 2024 Clinical Practice Guidelines on the Management of Asymptomatic Lower Limb Peripheral Arterial Disease and Intermittent Claudication. *Eur J Vasc Endovasc Surg* 2023.

4. Aboyans V, Ricco JB, Bartelink MEL, Bjorck M, Brodmann M, Cohnert T*, et al.* Editor's Choice - 2017 ESC Guidelines on the Diagnosis and Treatment of Peripheral Arterial Diseases, in collaboration with the European Society for Vascular Surgery (ESVS). *Eur J Vasc Endovasc Surg* 2018;**55**:305-68.

5. Frank U, Nikol S, Belch J, Boc V, Brodmann M, Carpentier PH*, et al.* ESVM Guideline on peripheral arterial disease. *Vasa* 2019;**48**:1-79.

6. Kotov A, Peters F, Debus ES, Zeller T, Heider P, Stavroulakis K*, et al.* The prospective GermanVasc cohort study. *Vasa* 2021;**50**:446-52.

7. Peters F, Behrendt CA. Limb Related Outcomes of Endovascular vs. Open Surgical Revascularisation in Patients with Peripheral Arterial Occlusive Disease: A Report from the Prospective GermanVasc Cohort Study. *Eur J Vasc Endovasc Surg* 2023;**66**:85-93.

8. Behrendt CA, Schwaneberg T, Hischke S, Muller T, Petersen T, Marschall U*, et al.* Data Privacy Compliant Validation of Health Insurance Claims Data: the IDOMENEO Approach. *Gesundheitswesen* 2020;**82**:S94-S100.

9. Behrendt CA, Sedrakyan A, Peters F, Kreutzburg T, Schermerhorn M, Bertges DJ*, et al.* Editor's Choice - Long Term Survival after Femoropopliteal Artery Revascularisation with Paclitaxel Coated Devices: A Propensity Score Matched Cohort Analysis. *Eur J Vasc Endovasc Surg* 2020;**59**:587-96.

10. Heidemann F, Kuchenbecker J, Peters F, Kotov A, Marschall U, L'Hoest H*, et al.* A health insurance claims analysis on the effect of female sex on long-term outcomes after peripheral endovascular interventions for symptomatic peripheral arterial occlusive disease. *J Vasc Surg* 2021;**74**:780-87 e7.

11. Heidemann F, Peters F, Kuchenbecker J, Kreutzburg T, Sedrakyan A, Marschall U*, et al.* Long Term Outcomes After Revascularisations Below the Knee with Paclitaxel Coated Devices: A Propensity Score Matched Cohort Analysis. *Eur J Vasc Endovasc Surg* 2020;**60**:549-58.

12. Kotov A, Heidemann F, Kuchenbecker J, Peters F, Marschall U, Acar L*, et al.* Sex Disparities in Long Term Outcomes After Open Surgery for Chronic Limb Threatening Ischaemia: A Propensity Score Matched Analysis of Health Insurance Claims. *Eur J Vasc Endovasc Surg* 2021;**61**:423-29.

13. Kreutzburg T, Peters F, Kuchenbecker J, Marschall U, Lee R, Kriston L*, et al.* Editor's Choice - The GermanVasc Score: A Pragmatic Risk Score Predicts Five Year Amputation Free Survival in Patients with Peripheral Arterial Occlusive Disease. *Eur J Vasc Endovasc Surg* 2021;**61**:248-56.

14. Behrendt CA, Kreutzburg T, Nordanstig J, Twine CP, Marschall U, Kakkos S*, et al.* The OAC(3)-PAD Risk Score Predicts Major Bleeding Events one Year after Hospitalisation for Peripheral Artery Disease. *Eur J Vasc Endovasc Surg* 2022;**63**:503-10.

15. Peters F, Behrendt CA. External Validation of the OAC3-PAD Risk Score to Predict Major Bleeding Events Using the Prospective GermanVasc Cohort Study. *Eur J Vasc Endovasc Surg* 2022.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

16. Lareyre F, Behrendt CA, Pradier C, Settembre N, Chaudhuri A, Fabre R*, et al.* Nationwide Study in France To Predict One Year Major Bleeding and Validate the OAC3-PAD Score in Patients Undergoing Revascularisation for Lower Extremity Arterial Disease. *Eur J Vasc Endovasc Surg* 2023;**66**:213-19.

17. Alushi K, Hinterseher I, Peters F, Rother U, Bischoff MS, Mylonas S*, et al.* Distribution of Mobile Health Applications amongst Patients with Symptomatic Peripheral Arterial Disease in Germany: A Cross-Sectional Survey Study. *J Clin Med* 2022;**11**:498.

18. Adegbola A, Behrendt CA, Zyriax BC, Windler E, Kreutzburg T. The impact of nutrition on the development and progression of peripheral artery disease: A systematic review. *Clin Nutr* 2022;**41**:49-70.

19. Wolbert L, Kreutzburg T, Zyriax BC, Adegbola A, Westenhöfer J, Jagemann B*, et al.* A cross-sectional survey study on the nutrition patterns of patients with peripheral artery disease. *Vasa* 2022.

20. Rosenberg Y, Görtz H, Rother U, Uhl C, Stavroulakis K, Pfeiffer M*, et al.* Empfehlungen zur konservativen Therapie und Sekundärprävention der peripheren arteriellen Verschlusskrankheit (PAVK): Eine evidenzbasierte Informationsbroschüre für Betroffene. *Gefässchirurgie* 2022;**27**:39-45.

21. Stock J, Petersen T, Behrendt C-A, Federrath H, Kreutzburg T. Privatsphärefreundliches maschinelles Lernen. *Informatik Spektrum* 2022;**45**:70-79.

22. Stock J, Petersen T, Behrendt C-A, Federrath H, Kreutzburg T. Privatsphärefreundliches maschinelles Lernen. *Informatik Spektrum* 2022;**45**:137-45.

23. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998;**36**:8-27.

24. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC*, et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;**43**:1130-9.

25. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama* 1982;**247**:2543-6.

26. Behrendt CA, Thomalla G, Rimmele DL, Petersen EL, Twerenbold R, Debus ES*, et al.* Editor's Choice - Prevalence of Peripheral Arterial Disease, Abdominal Aortic Aneurysm, and Risk Factors in the Hamburg City Health Study: A Cross Sectional Analysis. *Eur J Vasc Endovasc Surg* 2023;**65**:590-98.

27. Kreutzburg T, Peters F, Riess HC, Hischke S, Marschall U, Kriston L*, et al.* Editor's Choice - Comorbidity Patterns Among Patients with Peripheral Arterial Occlusive Disease in Germany: A Trend Analysis of Health Insurance Claims Data. *Eur J Vasc Endovasc Surg* 2020;**59**:59-66.

28. Peters F, Kreutzburg T, Kuchenbecker J, Marschall U, Remmel M, Dankhoff M*, et al.* Quality of care in surgical/interventional vascular medicine: what can routinely collected data from the insurance companies achieve? *Gefässchirurgie* 2020;**25**:19-28.

29. Behrendt C-A, Schwaneberg T, Hischke S, Müller T, Petersen T, Marschall U*, et al.* Data Privacy Compliant Validation of Health Insurance Claims Data – The IDOMENEO Approach. *Gesundheitswesen* 2019.

30. Carrasco-Ribelles LA, Llanes-Jurado J, Gallego-Moll C, Cabrera-Bean M, Monteagudo-Zaragoza M, Violán C*, et al.* Prediction models using artificial intelligence and longitudinal data

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

from electronic health records: a systematic methodological review. *J Am Med Inform Assoc* 2023;**30**:2072-82.

31. Kuchenbecker J, Peters F, Kreutzburg T, Marschall U, L'Hoest H, Behrendt CA. The Relationship Between Hospital Procedure Volume and Outcomes After Endovascular or Open Surgical Revascularisation for Peripheral Arterial Disease: An Analysis of Health Insurance Claims Data. *Eur J Vasc Endovasc Surg* 2022.

11. Anhang

Nicht zutreffend.

12. Anlagen

Anlage 1: Thesenpapier: Rechtliche Aspekte des Einsatzes von KI-Systemen

Anlage 2: Rechtsgutachten - Die Auswirkungen der Nutzung von KI-Software auf die ärztliche Haftung

Anlage 3: Ergebnisse der rechtswissenschaftlichen und informationstechnologischen Begleitung durch die Projektpartner der Universität Hamburg (gesperrt bis: 31.03.2025).

Anlage 4: Fragebogen zur Nutzerbefragung.

Anlage 5: Fragebogen zur Ernährung

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Thesenpapier: Rechtliche Aspekte des Einsatzes von KI-Systemen

MIN-Fakultät, FB Informatik, Arbeitsbereich Sicherheit in Verteilten Systemen (SVS)

Joshua Stock, Tom Petersen, Prof. Dr. Hannes Federrath
{joshua.stock, tom.petersen, hannes.federrath}@.uni-hamburg.de

9. November 2021

Inhaltsverzeichnis

1 Einleitung **4**

2 Grundlagen **5**
2.1 Lerntypen 5
2.2 Trainingsdaten und Merkmale 5
2.3 Verzerrung 6
2.4 Varianz, Unter- und Überanpassung 7
2.5 FairML 7

3 Verfahren **8**
3.1 Naive-Bayes-Klassifikation 8
3.2 k-Nearest-Neighbour 9
3.3 Support Vector Machines 10
3.4 Entscheidungsbäume und Random Forest 11
3.5 Lineare Regression 12
3.6 k-Means-Clustering 13
3.7 Principal Component Analysis (PCA) 14
3.8 Neuronale Netze 15

4 Erklärbares ML **17**

5 Angriffe auf maschinelle Lernverfahren **20**
5.1 Zugriff auf sensible Trainingsdaten 21
5.2 Deanonymisierung/Reidentifizierung 21
5.3 Model Inversion 22
5.4 Membership Inference 23
5.5 Property Inference 23
5.6 Model Extraction 24

6 Privacy-Preserving Machine Learning **24**
6.1 Pseudonymisierung und Anonymisierung 24
6.2 Differential Privacy 26
6.3 Federated Learning 29
6.4 Secure Multiparty Computation 31
6.5 Homomorphe Verschlüsselung 33
6.6 Trusted Execution Environments 34

7 Machine Learning: Anwendungen im Gesundheitswesen **35**
7.1 Kategorisierung von medizinischen ML-Anwendungen 36
7.1.1 Kategorisierungen in der Literatur 36
7.1.2 Trainingsdaten und Auswirkungen in der Inferenzphase 36
7.2 Regulierung 37
7.2.1 Behördliche Regulierung in den USA 37
7.2.2 Behördliche Regulierung in der EU 37
7.3 Fallgruppen 38

8 Thesen **41**
8.1 Rechtsgrundlage für die Verarbeitung 41
8.2 Datenschutz-Folgenabschätzung 42
8.3 Automatisierte Entscheidungen 43
8.4 Personenbezogene Daten: Angriffe 45
8.5 Rechtssichere technische und organisatorische Schutzmaßnahmen 47
8.6 Transparenz 49
8.7 Ausgaben von KI-Algorithmen als personenbezogene Daten 51
8.8 Recht auf Vergessenwerden 52
8.9 Recht auf Berichtigung 54

1 Einleitung

Der Einsatz von Künstlicher Intelligenz (KI) erfreut sich seit einigen Jahren immer größerer Beliebtheit, nicht zuletzt dank sinkender Kosten für Rechenleistung [DY14]. In diesem Dokument wird KI als Synonym zum Maschinellen Lernen (ML), einer Klasse selbstadaptiver Algorithmen, verstanden. Das adaptive Prozess der Algorithmen besteht aus einer Lern- bzw. Trainingsphase, die in der Regel dem Einsatz eines KI-Systems vorangeht. Nach einer erfolgreichen Trainingsphase können KI-Modelle dafür verwendet werden, anhand bisher unbekannter Daten Vorhersagen zu treffen. Je nach Einsatzgebiet können diese Vorhersagen unterschiedlicher Natur sein, beispielsweise:

- Verhaltensvorhersagen (Rückzahlungswahrscheinlichkeit von Krediten, Rückfälligkeit von Straftätern, ...),
- Mustererkennung (Objekterkennung auf Fotos, Spracherkennung, Gesichtserkennung, Charaktereigenschaften anhand von Gesichtszügen, ...),
- Medizinische Entscheidungshilfen (Datenanalyse einer Patientenakte, Krebsverdacht anhand von MR-/CT-Scans, ...).

Rechtssicherheit für die Verwendung von KI-Systemen ist nicht immer gegeben. Die Datenschutz-Grundverordnung (DSGVO) gibt zwar allgemeine Regeln für den Einsatz von Algorithmen vor (vor allem bezüglich automatisierter Entscheidungen in Art. 22 DSGVO), allerdings ist die Auslegung einzelner Passagen für KI-Systeme unklar. Das große Potential sinnvoller KI-Anwendungen kann sich erst durch die Schaffung von Rechtssicherheit abrufen lassen.

Zielsetzung Abschnitt 2 erklärt die allgemeinen Grundlagen maschineller Lernverfahren, wie beispielsweise gängige Lerntypen oder auftretende Verzerrungen beim Einsatz von ML-Verfahren. Abschnitt 3 stellt grundlegende Klassen maschineller Lernverfahren vor und gibt einen beispielhaften Einblick in häufig verwendete Techniken. In Abschnitt 4 wird das Prinzip des erklärbaren maschinellen Lernens dargestellt. Abschnitt 5 beschreibt verschiedene Angriffe auf ML-Modelle. Abschnitt 6 stellt schließlich verschiedene Verfahren für privatsphäreerhaltendes maschinelles Lernen dar. In Abschnitt 7 wird der Einsatz von ML-Anwendungen im Gesundheitswesen eingeordnet. Auf diesen Hintergrundkapiteln aufbauend werden in Abschnitt 8 Thesen zu rechtlichen Aspekten des ML aufgestellt. Diese sollen als interdisziplinäre Diskussionsgrundlage dienen und im Laufe des Diskurses fortwährend ergänzt, konkretisiert und mit Rechtsgrundlagen untermauert werden. Ziel ist die Prüfung der Thesen hinsichtlich ihrer Haltbarkeit und gegebenenfalls übersehener Implikationen. Falls Regelungsdefizite vorhanden sind, sollen diese explizit benannt werden.

2 Grundlagen

Dieser Abschnitt erklärt Überkategorien für Algorithmen des maschinellen Lernens und stellt grundlegende Begriffe vor.

2.1 Lerntypen

Es gibt zahlreiche Algorithmen, die alle unter dem Begriff „Maschinelles Lernen“ (ML) zusammengefasst werden können. Sie lassen sich dabei grob in die drei Kategorien (bzw. Paradigmen) überwachtes Lernen, unüberwachtes Lernen und bestärkendes Lernen (engl. *supervised*, *unsupervised* und *reinforcement learning*) einteilen [Alp19].

Mit der ersten Kategorie, dem **überwachten Lernen**, können Klassifikationsprobleme (z.B. Mustererkennung, Prädiktion oder Ausreißererkennung) sowie Regressionsprobleme (Vorhersage von Werten) gelöst werden [Alp19]. Algorithmen, die dabei zum Einsatz kommen, müssen auf eine Dateneingabe X eine (möglichst korrekte) Ausgabe Y liefern, die in der Regel Wahrscheinlichkeit-basiert ist. ML-Algorithmen des überwachten Lernens erlernen diese Abbildung von X zu Y in einer Trainingsphase. Im überwachten Lernen sind die in dieser Phase genutzten Daten stets mit ihrem jeweiligen Zielwert $y \in Y$ annotiert.

In der Kategorie des **unüberwachten Lernens** hingegen sind die Ausgabewerte Y zu den Trainingsdaten nicht von vornherein bekannt [Alp19]. Ein klassisches Anwendungsbeispiel sind Clusteranalysen, die Regelmäßigkeiten in den Eingabewerten erkennen und ähnliche Datensätze in Clustern zusammenfassen. Die Cluster werden dabei dynamisch in Abhängigkeit von der jeweiligen Datenbeschaffenheit erstellt. Weitere mögliche Anwendungen für unüberwachtes Lernen sind in der Dimensionalitätsreduktion zu finden, die beispielsweise für die Komprimierung genutzt wird.

Die dritte Kategorie, **bestärkendes Lernen**, ermöglicht das Lernen von Verhaltensweisen bzw. (Abfolgen von) Aktionen [Alp19]. Ein beliebtes Anwendungsfeld sind Brettspiele. Charakteristisch für die Praktikabilität des bestärkenden Lernens ist hierbei, dass nicht durch jede einzelne Aktion (bzw. durch jeden Spielzug) unmittelbar ein positiver Effekt eintreten muss, sondern die langfristigen Strategien der Akteure den Spielausgang bestimmen. In der Trainingsphase erlernt ein ML-Algorithmus hierbei durch das wiederholte, experimentelle Anwenden verschiedener Aktionssequenzen und die gleichzeitige Beobachtung der jeweiligen Spielausgänge, welche Strategien am wirksamsten sind.

Neben diesen drei Grundparadigmen existieren Mischformen, kleinere Kategorien und Unterkategorien, die den Umfang dieses Abschnitts übersteigen. Für die Umsetzung der jeweiligen Lernformen stehen eine Vielzahl an Verfahren und Algorithmen bereit, die oft auch für mehrere Kategorien eingesetzt werden können.

2.2 Trainingsdaten und Merkmale

Für überwachtes und nicht-überwachtes Lernen, die hier im Fokus stehen, wird ein sogenannter Trainingsdatensatz benötigt. Während der Lern- oder auch Trainingsphase wird ein Modell auf existierenden Trainingsdaten trainiert, um das gewünschte Verhalten zu erlernen. In der Inferenzphase wird ein so trainiertes Modell dazu genutzt, das gewünschte Verhalten auf neue Daten anzuwenden. Zum Training wird ein Trainingsdatensatz $x = (x^{(1)}, \ldots, x^{(l)})$ verwendet, der aus l Datenpunkten $x^{(i)}$ besteht. Jeder dieser Datenpunkte $x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$ wird durch n Merkmale (engl. *Features*) beschrieben. Hierbei kann es sich je nach Szenario um verschiedene Merkmale

handeln, wie beispielsweise die Medikamentdosierung oder den Blutdruck während einer Patientenbehandlung, die einzelnen Pixel eines CT-Scans oder auch die Freitextbeschreibung eines behandelnden Arztes. Die Gesamtheit aller möglichen Merkmale bildet den Merkmalsraum (engl. *Featurespace*).

Ein Beispiel zur Verdeutlichung: Es soll ein Modell, das das Risiko einer notwendigen Amputation für PAVK-Erkrankte vorhersagen soll, mithilfe eines überwachten Lernverfahrens trainiert werden. In der Trainingsphase wird ein Trainingsdatensatz bestehend aus einer Vielzahl von Datenpunkten, jeweils einen Patienten beschreibend, genutzt. Jeder dieser Datenpunkte besteht aus den gleichen Merkmalen, wie bspw. dem Patientenalter oder der Informationen über den Alkoholkonsum des Patienten sowie der Annotation mit dem Zielwert, in diesem Beispiel also der Angabe, ob eine Amputation erfolgte oder nicht. Ein mit diesen Daten trainiertes Modell kann anschließend in der Inferenzphase dazu genutzt werden, für neue Datenpunkte vorherzusagen, wie hoch das Risiko für eine zukünftig notwendige Amputation ist.

2.3 Verzerrung

Verzerrung (manchmal auch systematische Abweichung oder systematischer Fehler, engl. *Bias*) beschreibt die Abweichung eines durch ein ML-Verfahren vorhergesagten Wertes vom tatsächlichen Wert. Eines der bekanntesten Beispiele für Verzerrungen in ML-Verfahren wurde 2016 in dem Artikel *Machine Bias* der Organisation ProPublica veröffentlicht, in dem über Bias in der zur Beurteilung von Straftätern verwendeten Software COMPAS berichtet wurde [Ang+16]. Der in COMPAS eingesetzte Algorithmus besaß signifikant unterschiedliche Falsch-Positiv-Fehlerraten für das erneute Begehen von Straftaten für verschiedene Bevölkerungsgruppen. Die Wahrscheinlichkeit eines Rückfalls für afroamerikanische Straftäter wurde im Gegensatz zu Mitgliedern anderer Bevölkerungsgruppen systematisch überschätzt. Verzerrung kann aus vielfältigen Gründen auftreten [Meh+21], beispielsweise:

- *Sample Bias* entsteht durch eine nicht der Realverteilung entsprechenden Auswahl von Trainingsdaten. Ein Beispiel ist das Trainieren eines Gesichtserkenungsverfahrens auf lediglich hellhäutigen Gesichtern, das in der Praxis an der Erkennung dunkelhäutiger Gesichter scheitert [BG18].

- *Exclusion Bias* beschreibt eine Verzerrung, die durch den Ausschluss von relevanten Merkmalen vor der Trainingsphase bedingt ist. Häufig wird durch den Entwickler eines ML-Systems aus Performancegründen eine Vorauswahl relevanter Features getroffen, die die Vorhersagekraft des Systems direkt beeinflusst. Wird beispielsweise in einem Verfahren zur Mietpreisvorhersage der Standort des Objekts nicht berücksichtigt, wird das System Faktoren wie Flughafennähe nicht vorhersagen können.

- *Measurement Bias* bezeichnet eine Verzerrung, die durch systematisch unterschiedliche Trainings- und Produktivdaten hervorgerufen wird. Ein Beispiel ist die Nutzung unterschiedlicher CT-Scanner während des Trainings und dem produktiven Einsatz eines ML-Verfahrens zur Bilderkennung.

- *Algorithmic Bias* beschreibt Verzerrungen, die durch eine ungeschickte Auswahl von ML-Verfahren oder deren Parametern entstehen. Beispielsweise kann der Einsatz einer einfachen Regression, die lediglich lineare Zusammenhänge abbildet, in komplexen Szenarien zu großen Verzerrungen führen.

Auch bei der Anwendung von ML-Verfahren im medizinischen Bereich besteht die Gefahr verschiedenster Verzerrungen [Gia+18].

2.4 Varianz, Unter- und Überanpassung

Zusätzlich zur Verzerrung kann auch die Varianz eines Modells betrachtet werden. Varianz kann als die Empfindlichkeit des Modells gegenüber kleinen Schwankungen in den Trainingsdaten angesehen werden. Im Bereich des überwachten Lernens tritt das sogenannte Verzerrungs-Varianz-Dilemma auf. Verfahren können entweder die Verzerrung oder die Varianz minimieren, aber nicht beides.

In der Praxis führt dieses Dilemma dazu, dass es beim Training von Modellen zu *Unteranpassung* oder *Überanpassung* kommen kann. Unteranpassung (engl. *Underfitting*) beschreibt ein Modell mit großer Verzerrung, dessen Vorhersagen den echten statistischen Zusammenhang nicht abbilden. Überanpassung (engl. *Overfitting*) beschreibt ein Modell mit großer Varianz, das zu sehr auf den Trainingsdaten beruht und beispielsweise vorliegende Messfehler oder Rauschen mit einbezieht. Eine Visualisierung dieser Konzepte für Regressions- und Klassifikationsprobleme ist in Abbildung 1 zu finden.

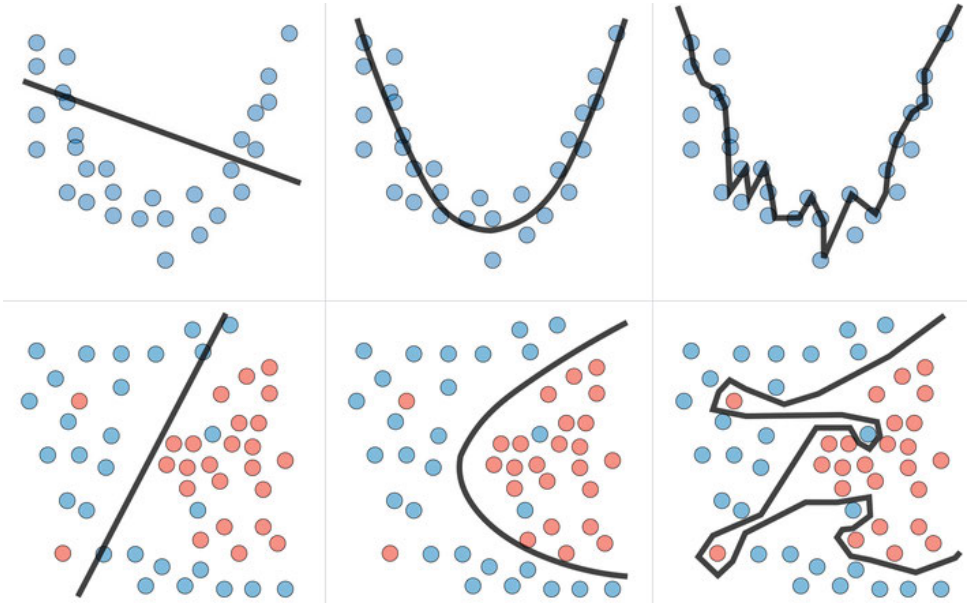

Abbildung 1: Visualisierung von Unteranpassung (links), einem guten Modell (Mitte) und Überanpassung (rechts) für ein Regressionsproblem (oben) und ein Klassifkationsproblem (unten). Die Abbildung stammt aus dem *Machine Learning Cheat Sheet* von Afshine and Shervine Amidi (https://github.com/afshinea/stanford-cs-229-machine-learning).

2.5 FairML

Um Verzerrungen (Abschnitt 2.3) vorzubeugen und entgegenzuwirken, rückt der Begriff der *Fairness* zunehmend in den Fokus von Forschung und Praxis. Fairness (oft auch unter dem Begriff *Ethical AI* geführt) bezeichnet dabei vor allem die Eigenschaft von ML-Algorithmen, niemanden zu diskrimieren. Da Diskriminierungen in vielen Bereichen des Lebens an der Tagesordnung steht, schlägt sich dies auch oft in den für das Training von ML-Algorithmen verwendeten Datensätzen nieder. Das Erreichen von Fairness ist daher nicht trivial.

Einerseits haben sich drei mathematische Ansätze für die Entwicklung fairer ML-Algorithmen etabliert [CG18]: Das systematische Ausschließen von sensiblen Attributen wie Geschlecht oder ethnischer Herkunft in Lernverfahren (*anti-classification*), das künstliche Ausgleichen von unterschiedlichen Falsch-Positiv-Fehlerraten für jeweilige Gruppen in den Trainingsdaten (*classification parity*, siehe Negativbeispiel in Abschnitt 2.3) und das Sicherstellen von statistische

Unabhängigkeit zwischen Ausgaben und geschützten Attributen (*calibration*). Jene mathematischen Ansätze bedürfen allerdings viel Feingefühl in der Umsetzung, da sie ansonsten durch die systematische Andersbehandlung verschiedener Gruppen Verzerrungen unter Umständen weiter befeuern können, anstatt diese zu beheben [CG18].

Es existieren zahlreiche Frameworks wie *AI Fairness 360* von IBM oder Aequitas [Sal+18], die Modellentwickler dabei unterstützen können, ihre Algorithmen unter verschiedenen Fairness-Aspekten zu beleuchten und zu verbessern. Eine umfassende Übersicht über Fairness in ML bietet [Meh+21].

3 Verfahren

In diesem Abschnitt werden einige gängige ML-Verfahren kurz dargestellt, die die Grundlagen für die Anwendung von maschinellem Lernen im Bereich der Medizin darstellen.

3.1 Naive-Bayes-Klassifikation

Die Naive-Bayes-Klassifikation ist ein Verfahren, das der überwachten Klassifikation dient und auf dem Satz von Bayes beruht. Eingangsdaten, die aus mehreren Features bestehen, werden einer Klasse C_k aus einer festen Menge an Klassen $C = \{C_1, \ldots, C_m\}$ zugewiesen. Es wird die vereinfachende Annahme der Unabhängigkeit der Features der Eingangsdaten getroffen.

Für einen aus n Features bestehenden Datensatz $x = (x_1, \ldots, x_n)$ berechnet der Naive-Bayes-Klassifikator das wahrscheinlichste Klassenlabel $\hat{y} \in C$ als

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \, p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Die Klassenwahrscheinlichkeiten $p(C_k)$ können als gleichwahrscheinlich angenommen werden oder aus der Verteilung innerhalb der Trainingsdaten berechnet werden. Die Wahrscheinlichkeitsverteilung für Features innerhalb einer Klasse $p(x_i \mid C_k)$ können basierend auf der Annahme einer Verteilung aus den Trainingsdaten berechnet werden. Abhängig von den Featuredaten kann so beispielsweise eine Normalverteilung mit aus den Trainingsdaten abgeleiteten Parametern verwendet werden (*Gaussian Naive Bayes*).

Aufgrund seiner Einfachheit funktioniert das Trainieren des Klassifizierers schnell und benötigt eine geringe Menge an Trainingsdaten. Trotz dieser Einfachheit und obwohl die Unabhängigkeitsannahme der Eingangsfeatures häufig nicht haltbar ist, liefert der Naive-Bayes-Klassifier in der Praxis häufig erstaunlich gute Klassifikationsergebniss [Ris01].

3.2 k-Nearest-Neighbour

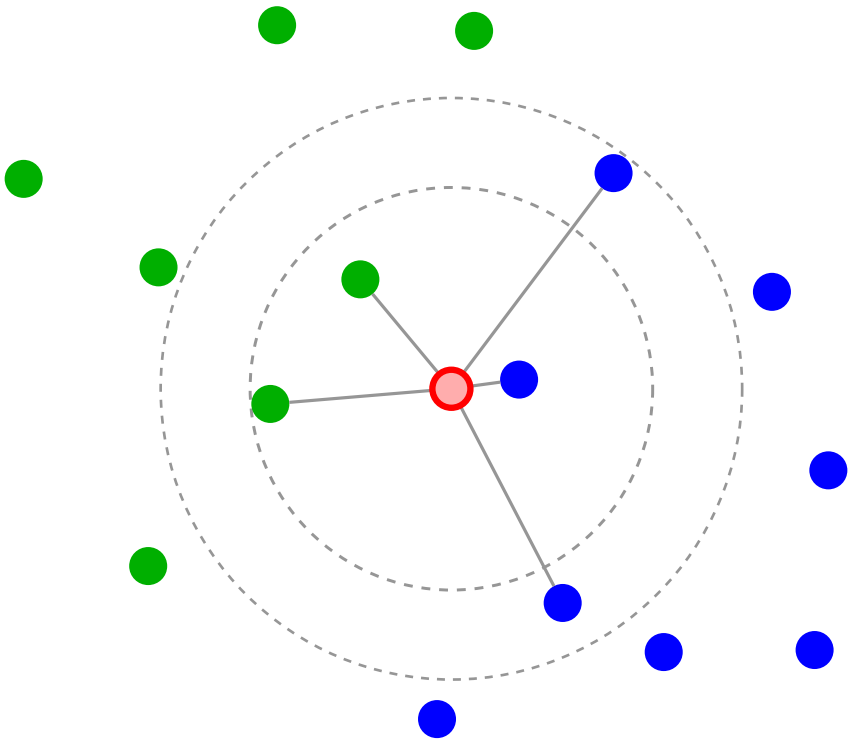

Abbildung 2: k-Nearest-Neighbour-Klassifikation für die Werte $k = 3$ und $k = 5$. Für die unterschiedlichen Werte ergeben sich für die Klassifizierung des rot umrandeten Zieldatensatzes unterschiedliche Ergebnisklassen: Für $k = 3$ wird das grüne Klassenlabel vergeben, für $k = 5$ wird der Datensatz als der blauen Klasse zugehörig klassifiziert.

Der *k-Nearest-Neighbour*-Algorithmus (kNN) kann zur überwachten Klassifikation oder Regression eingesetzt werden. Die Grundidee besteht darin, für einen Datenpunkt die k ähnlichsten Datenpunkte (Nachbarn) innerhalb der Trainingsdaten zu finden und als Ergebnis die häufigste Klasse (Klassifikation) oder den Durchschnitt der Zielwerte (Regression) dieser Nachbarn zu verwenden. Abbildung 2 stellt das Verfahren für die Klassifikation dar.

In der Praxis sind hierbei verschiedene Entscheidungen zu treffen. Neben der Wahl eines geeigneten Parameters k und eines geeignetes Maßes zur Bestimmung des Abstands zweier Datensätze entscheidet insbesondere auch die Auswahl geeigneter Features über die Qualität der Ergebnisse. Um eine bessere Performance zu erreichen, können vor der eigentlichen Ausführung auch wenige aussagekräftige Vertreter aus den Trainingsdaten ausgewählt werden, die dann zur Bestimmung der Nachbarn genutzt werden. Auf diese Weise müssen weniger Distanzen berechnet werden. Alternativ hierzu steht eine Vielzahl von Verfahren zur Verfügung, die durch den geschickten Einsatz von Datenstrukturen oder Heuristiken ein schnelleres Finden der Nachbarn erlauben [Ben75].

3.3 Support Vector Machines

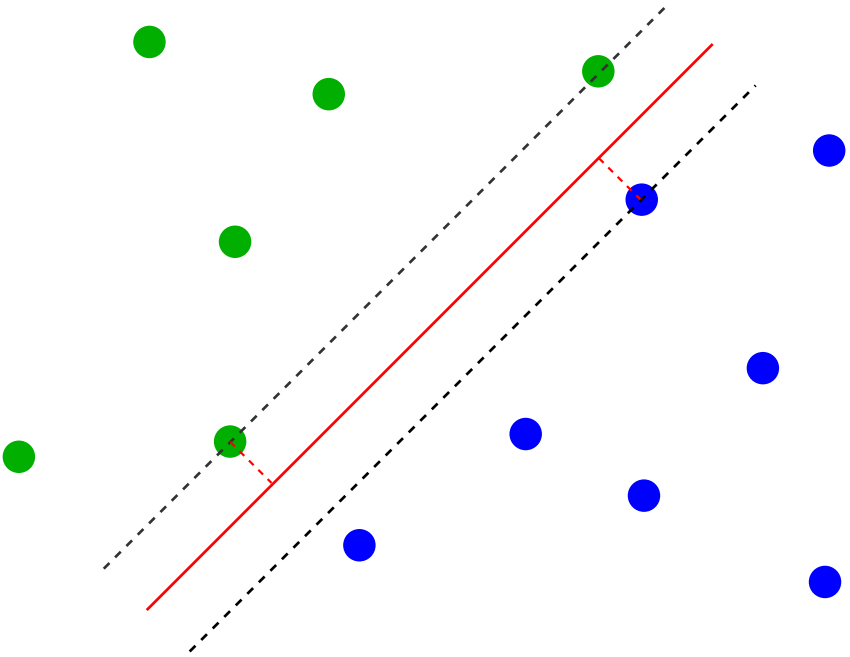

Abbildung 3: Darstellung einer einfachen Klassifikation durch eine Support Vector Machine mit den dazugehörigen Support Vectors.

Support Vector Machines (SVMs, dt. *Stützvektormaschinen*) werden zur überwachten Klassifikation eingesetzt [CV95]. Grundsätzlich sind SVMs nur für binäre Klassifikationsaufgaben (mit zwei möglichen Klassen) geeignet. Die grundlegende Idee besteht darin, in der Trainingsphase eine bestmögliche Trennung der Trainingsdaten in die beiden Klassen durch Hyperebenen im Merkmalsraum zu finden. Hierzu wird eine Hyperebene ermittelt, die eine maximale Distanz zu den ihr am nächsten liegenden Trainingsdatensätzen (*Support Vectors*) der entsprechenden Klassen erreicht. Durch Maximierung der Distanz sollen gute Ergebnisse für spätere Klassifikationen erreicht werden, selbst wenn in den Trainingsdaten unähnliche Daten verwendet werden. Abbildung 3 zeigt ein einfaches Beispiel inklusive der trennenden Hyperebene (in diesem Fall eine Gerade) und der *Stützvektoren* (engl. *Support Vectors*). Da viele Klassifizierungsprobleme nicht linear lösbar sind, können sich SVMs des sogenannten *Kernel-Tricks* bedienen, der die Daten in höherdimensionale Daten abbildet, für die eine lineare Trennung möglich ist.

Eine Erweiterung für die Klassifikation mit mehr als zwei Klassen besteht darin, mehrere Einzelprobleme als *Ist-in-Klasse vs. Ist-nicht-in-Klasse* zu betrachten und die Einzelergebnisse geeignet zu verknüpfen. Es existieren auch Erweiterungen zur Nutzung des Prinzips für die Regressionsanalyse [Dru+96].

Ein Vorteil von SVMs besteht darin, dass sie auch für hochdimensionale Daten, also Daten, die viele Merkmale besitzen, und wenige Trainingsdatensätze gute Ergebnisse liefern können.

3.4 Entscheidungsbäume und Random Forest

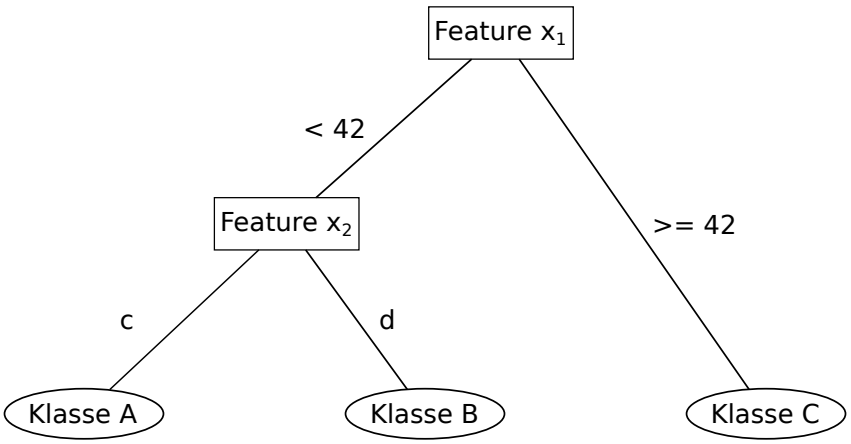

Abbildung 4: Ein einfacher Entscheidungsbaum, der Datensätze basierend auf den zwei Features x_1 und x_2 in drei Klassen A, B und C einordnet.

Entscheidungsbäume dienen der überwachten Klassifikation (*classification trees*) und Regression (*regression trees*) [Bre+84]. In Abbildung 4 ist ein einfacher Entscheidungsbaum für die Klassifikation abgebildet. Während der Trainingsphase wird der Entscheidungsbaum in einem *top-down*-Ansatz vom Wurzelknoten aus aufgebaut. Hierzu werden rekursiv Knoten anhand eines Merkmales partitioniert, bis die entstehenden Blattknoten ausreichend homogen in Bezug auf die Zielklasse oder den Zielwert sind. Datensätze werden während der Inferenzphase ausgehend vom Wurzelknoten eines Baumes bis zu einem Blattknoten bewertet. Hierzu werden an den Knoten des Baumes Entscheidungen in Bezug auf ein Feature getroffen und so ein Pfad bis zu einem Blattknoten gewählt. Der Blattknoten beschreibt die Klasse oder den Zielwert für den Datensatz. Da ein optimaler Entscheidungsbaumes schwierig zu berechnen ist[1], werden in der Praxis heuristische Ansätze wie CART verwendet [Bre+84].

Entscheidungsbäume bilden ein leicht zu verstehendes und gut zu visualisierendes Verfahren, das dadurch auch leicht zu interpretieren ist [Knu21]. In ihrer einfachsten Form neigen sie jedoch insbesondere bei der Betrachtung von vielen Features zu *Overfitting*. Abhilfe kann der Einsatz mehrerer, separat trainierter Entscheidungsbäume schaffen, auf deren Ergebnissen eine Mehrheitsentscheidung getroffen wird. Dieses Verfahren wird *Random Forest* genannt. Das Training der verschiedenen Bäume erfolgt hierbei auf einer zufälligen Menge der Trainingsdaten und mit einer Zufallsauswahl relevanter Merkmale für die Partitionierungen. Hierdurch werden unterschiedliche Entscheidungsbäume erreicht. Die Bewertung eines Datensatzes in der Inferenzphase erfolgt durch separate Bewertung in jedem Baum. Das finale Ergebnis für den Datensatz bildet die Klasse, die am häufigsten von den einzelnen Entscheidungsbäumen als Resultat berechnet wurde.

[1] Genauer liegt es in der Klasse der NP-vollständigen Probleme, einem Konzept aus der Komplexitätstheorie. Es wird vermutet, dass für diese Art von Problemen keine effizienten, in polynomieller Zeit berechenbaren Lösungsverfahren existieren.

3.5 Lineare Regression

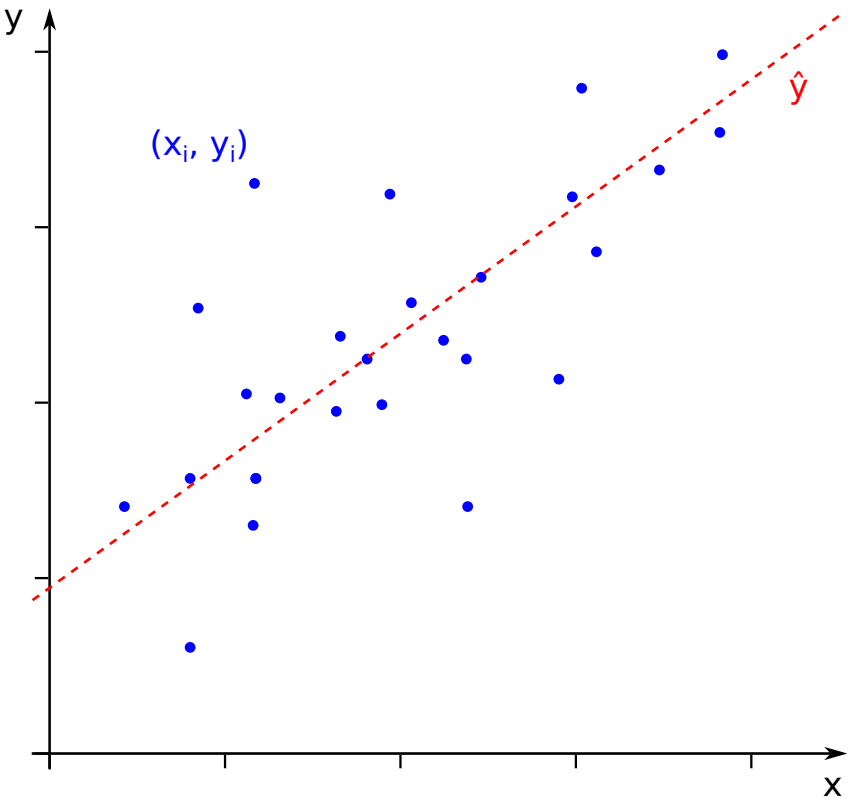

Abbildung 5: Beispiel einer einfachen linearen Regression. Die blauen Punkte stellen die einzelnen Trainingsdaten dar, die rote Linie die durch die Trainingsdaten bedingte Vorhersage.

Lineare Regressionsverfahren dienen der überwachten Regression von Trainingsdaten. Die Verfahren wählen dabei Linearparameter β, sodass die für einen Trainingsdatensatz $x_i = (x_1^{(i)}, \ldots, x_n^{(i)})$ beobachteten Werte y_i und die durch die Linearfunktion berechneten Werte $\hat{y}_i = \beta_0 + \beta_1 \cdot x_1^{(i)} + \cdots + \beta_n \cdot x_n^{(i)}$ möglichst wenig voneinander abweichen. Eine einfache lineare Regression ist in Abbildung 5 dargestellt.

Eine mögliche Variante ist die sogenannte *Methode der kleinsten Quadrate*, bei der für die Abweichung die Summe der Fehlerquadrate betrachtet wird. Es entsteht ein Minimierungsproblem:

$$\min_{\beta} \sum_{i=1}^{l} (\hat{y}_i - y_i)^2$$

Da dieses Verfahren in seiner einfachsten Form zu *Overfitting* neigt, können zusätzliche Nebenbedingungen, sogenannte Strafterme, hinzugefügt werden. Man spricht hierbei auch von *Regularisierung*. Beispiele hierfür sind LASSO- und Ridge-Regularisierung [Tib96; HK70].

3.6 k-Means-Clustering

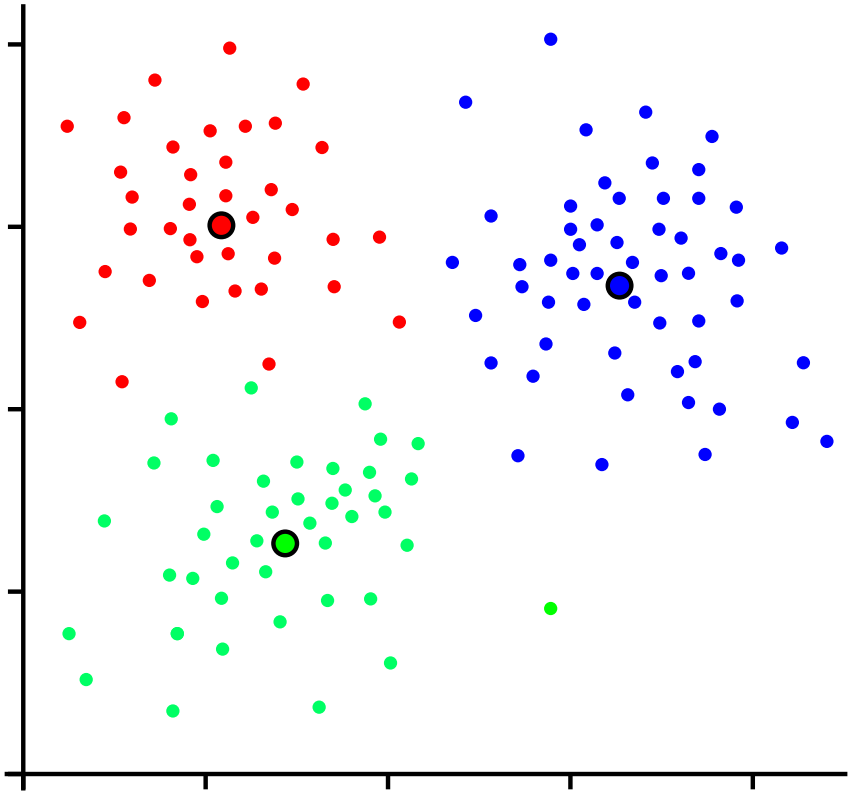

Abbildung 6: Beispiel eines k-Means-Ergebnisses für $k = 3$. Die Cluster-Mittelwerte werden durch die gerahmten Datenpunkte dargestellt.

Der k-Means-Algorithmus wird zum unüberwachten Clustering von Daten verwendet [Mac+67]. Ziel ist es, Daten derartig in k Cluster $S = \{S_1, \ldots, S_k\}$ mit zugehörigen Cluster-Mittelwerten μ_i (auch Schwerpunkte oder Zentroide genannt) zu partitionieren, dass die Summe aller Abweichungen von diesen Mittelwerten minimal ist. Abbildung 6 stellt ein mögliches Resultat des Algorithmus für drei Cluster dar. Es ergibt sich das Optimierungsproblem der Minimierung von

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

über die Zuweisung der Daten x_j zu den Clustern S_i. Das Finden einer optimalen Lösung für dieses Problem ist schwierig[2], es existieren jedoch viele heuristische Verfahren. Ein Beispiel für ein solches ist Lloyd's Algorithmus. Hierbei werden initial k Mittelwerte zufälllig gewählt (beispielsweise schlicht k zufällige Datensätze). Anschließend wird jeder Datensatz jeweils dem am nächsten liegenden Mittelwert zugewiesen und es werden für die so entstehenden Cluster neue Mittelwerte berechnet. Dieser Vorgang wird so lange wiederholt, bis sich die Zuweisung von Datensätzen zu Clustern nicht mehr ändert.

[2]Genauer handelt es sich um ein NP-schweres Problem, siehe Fußnote 1.

3.7 Principal Component Analysis (PCA)

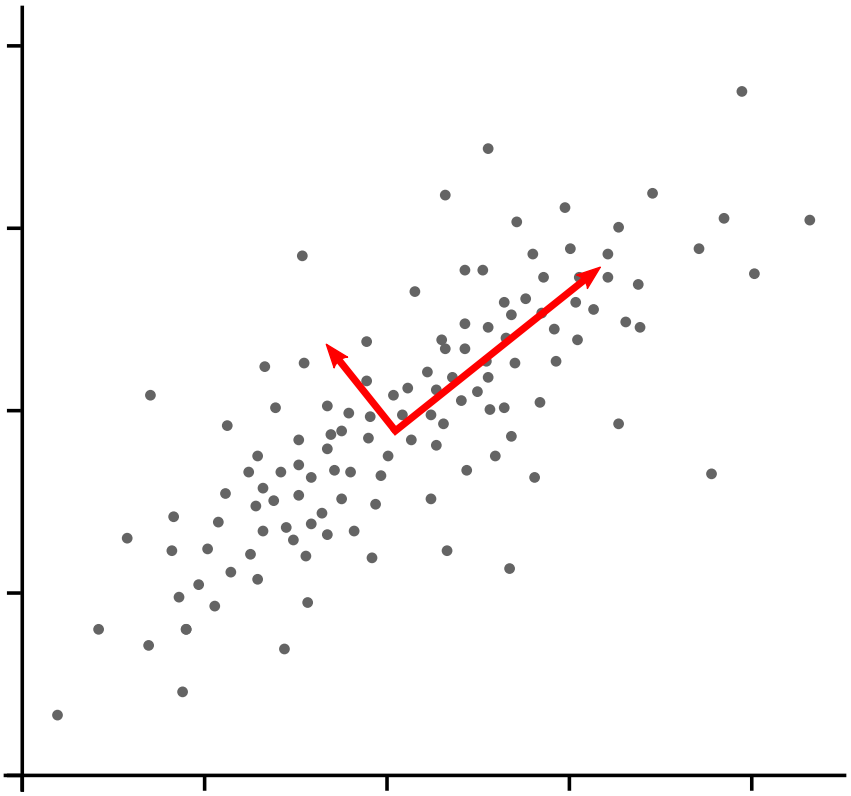

Abbildung 7: Ein Beispiel für das Resultat einer *Principal Component Analysis*. Das Koordinatensystem stellt den ursprünglichen 2-dimensionalen Merkmalsraum dar. Die Pfeile stellen das durch die PCA ermittelte neue Koordinatensystem dar, dessen Basisvektoren jeweils in Richtung der größten Varianz der Trainingsdaten liegen.

Dimensionsreduktion beschreibt die Abbildung von Daten in einen Raum mit weniger Dimensionen möglichst unter Beibehaltung charakteristischer Eigenschaften der Ursprungsdaten. Für die Dimensionsreduktion von Daten lassen sich zwei Klassen von Verfahren unterscheiden: *Feature selection*, die lediglich jene k der n Merkmale mit der größten Varianz auswählt, und *Feature Reduction*, bei der aus n Merkmalen k neue Merkmale berechnet werden, die die größtmögliche Varianz innerhalb der Daten darstellen. Die Grundannahme dieser Verfahren besteht darin, dass die Merkmale mit größter Varianz auch die meisten Informationen transportieren, sodass eine Datenanalyse auch nach der Dimensionsreduktion aussagekräftige Ergebnisse liefern kann.

Principal Component Analysis (PCA, dt. Hauptkomponentenanalyse) ist eine lineare Methode zur *Feature extraction*. Das Verfahren berechnet für einen Datensatz einen neuen *Merkmalsraum*, in dem die orthogonalen Basisvektoren jeweils in die Richtung zeigen, die die größte Varianz in den Ursprungsdaten enthält. Ein einfaches Beispiel ist in Abbildung 7 dargestellt. Hier werden Trainingsdaten in einem zweidimensionalen Merkmalsraum auf einen neuen (ebenfalls zweidimensionalen) Merkmalsraum abgebildet. Die Basisvektoren dieses neuen Merkmalsraumes bilden die größte Varianz innerhalb der Trainingsdaten ab. In der Praxis würde die Abbildung von einem höherdimensionalen Merkmalsraum in einen mit weniger Dimensionen erfolgen, um unter Beibehaltung größtmöglicher Varianz den Zweck der Dimensionsreduktion zu erfüllen.

3.8 Neuronale Netze

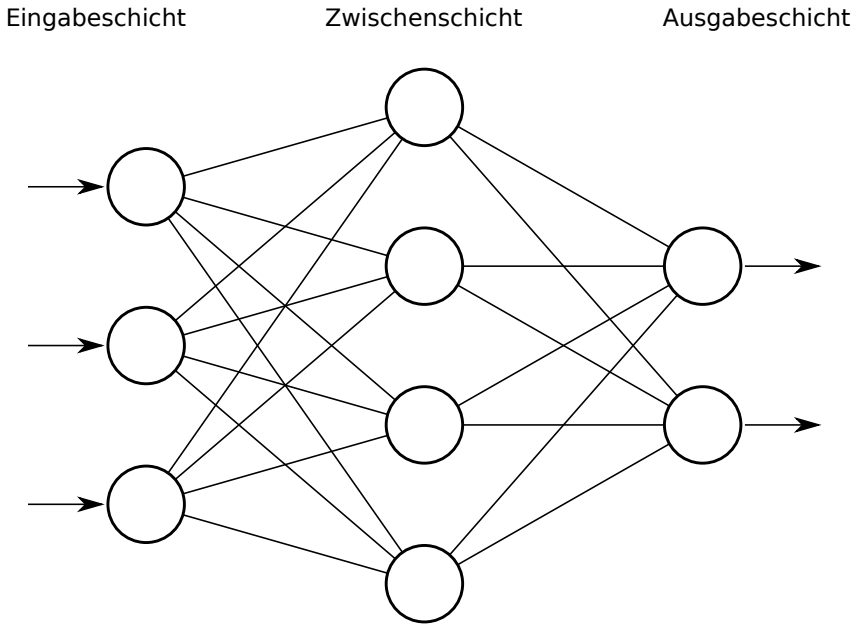

Abbildung 8: Schematische Darstellung der Schichten eines neuronalen Netzes.

Neuronale Netze in unterschiedlichen Ausprägungen haben eine breite Anwendbarkeit in verschiedenen Aufgaben des maschinellen Lernens. Neben überwachter Klassifikation und Regression, unüberwachtem Clustering und der Detektion von Anomalien in Daten können mit ihrem Einsatz auch (in echtem Wortsinne) übermenschliche Leistungen in klassischen Brettspielen wie Schach und Go erzielt werden.

Die grundlegende Idee ist inspiriert vom Aufbau und von den Vorgängen im menschlichen Gehirn. Ihre kleinsten Bestandteile bilden sogenannte Neuronen, die üblicherweise in Schichten angeordnet und untereinander durch Kanten verbunden sind. Ein Netz besteht wie in Abbildung 8 dargestellt aus einer Eingabeschicht (engl. *input layer*) (beispielsweise bestehend aus einem Neuron pro Pixel eines Eingabebildes), einer Menge von Zwischenschichten (engl. *hidden layers*) und einer Ausgabeschicht (engl. *output layer*), die wiederum beispielsweise ein Neuron pro möglicher Klasse in einem Klassifizierungsproblem enthält. Werden viele Zwischenschichten verwendet, so wird auch von *Deep Neural Networks* oder allgemeiner von *Deep Learning* gesprochen [Sch15]. Ähnlich zu den Vorgängen im Gehirn führen verschiedene Eingaben zu unterschiedlich starken Aktivierungen der Neuronen. Abhängig von ihrer Aktivierungsintensität geben die Neuronen über ihre Kantenverbindungen Signale an die nächste Schicht weiter. Das Ergebnis kann letztlich aus der Aktivierung der Ausgabeneuronen in der letzten Schicht abgelesen werden.

Im Detail wird die jeweilige Aktivierung folgendermaßen berechnet: Ein Neuron empfängt zunächst eine Menge von Eingaben I und berechnet daraus mithilfe einer nicht-linearen Aktivierungsfunktion Φ über die Summe aller Eingaben seine Ausgabe O. Beliebte Aktivierungsfunktionen sind *ReLU* („Rectified Linear Unit“) und *Sigmoid*. Erstere schneidet mittels $\Phi_{\text{ReLU}}(x) = max(0,x)$ effektiv alle Werte < 0 ab, sodass die Funktionsausgabe immer positiv ist. Die Sigmoid-Funktion $\Phi_{\text{sig}}(x) = \frac{1}{1+e^{-x}}$ bildet alle Eingaben auf einem Wert zwischen 0 und 1 ab. Die beiden Funktionsgraphen sind in der Abbildung 9 dargestellt.

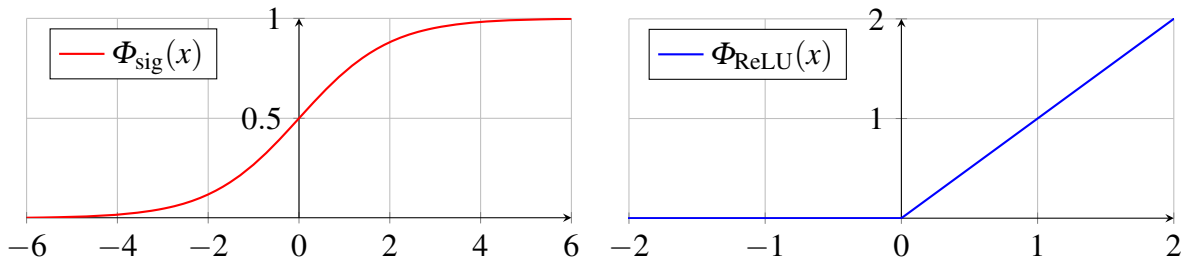

Abbildung 9: Funktionsgraphen der Aktivierungsfunktionen Sigmoid $\Phi_{\text{sig}}(x) = \frac{1}{1+e^{-x}}$ und ReLU $\Phi_{\text{ReLU}}(x) = max(0,x)$, die innerhalb von Neuronen zum Einsatz kommen können.

Die Eingaben für einzelne Neuronen I werden jeweils mit Gewichten w (von engl. *weights*) gewichtet und zusätzlich mit einem sogenannten Bias-Term b addiert:

$$O = \Phi(b + \sum_{i=1}^{k} w_i \cdot I_i)$$

Zu Beginn des Trainingsprozesses eines neuronalen Netzes werden die Gewichte meist mit Zufallswerten initialisiert. Die Trainingsphase besteht nun darin, dass das Netz mit Trainingsdaten gespeist wird und unter Beobachtung der Ausgabewerte die Modellparameter (Kantengewichte und Bias-Terme) angepasst werden. Als Ziel der Anpassung dient die Minimierung des beobachteten Fehlers, beispielsweise der Abweichung des Ergebnisses zur echten Klasse des Trainingsdatensatzes in einem überwachten Lernprozess. Dieser Prozess wird *Back-Propagation* genannt, da der Fehler von der letzten Schicht über die Zwischenschichten bis zu den Kantengewichten der Eingabeschicht rückwärts durch das Netz propagiert wird und dabei die entsprechenden Gewichte angepasst werden.

Es gibt zahlreiche Strategien für einen effizienten und effektiven Trainingsprozess, die je nach Aufgabengebiet, Datenbeschaffenheit und Kontext variieren. Beliebt ist beispielsweise die Aufteilung der Trainingsdaten in sogenannte *(Mini-)Batches*, um die Modellparameter nicht nach jedem einzelnen Datum, sondern erst nach Beobachtung des Netzes bei der Eingabe eines ganzen *Batches* anzupassen. Zusätzlich werden die Trainingsdaten wiederholt in mehreren Durchläufen (engl. *epochs*) verarbeitet, in der Regel einmal pro Durchlauf. Dabei kann die Zusammenstellung der *Batches* auch nach jedem Durchlauf variiert werden, um die Auswirkungen unausgeglichener *Batch*-Zusammenstellungen zu minimieren. Auch die Lernrate (engl. *learning rate*), also das Flexibilitätsmaß der Modellparameter, mit dem die beobachteten Fehler korrigiert werden, kann im Verlauf des Lernprozesses geändert werden. In der Regel wird mit einer höheren Lernrate begonnen, um die initialen Zufallswerte der Modellparameter schnell zu korrigieren. Nach einigen Durchläufen kann die Lernrate reduziert werden, um das Modell schrittweise gegen ein lokales Optimum (hinsichtlich der Fehlerminimierung) konvergieren zu lassen.

Neben der Lernrate, der Anzahl von *Batches* und Trainingsdurchläufen muss auch der Aufbau eines Netzes (die Anzahl der Neuronenschichten, die Anzahl der Neuronen pro Schicht und weitere Eigenschaften) durch den Entwickler beim Entwurf des Netzes festgelegt werden – diese Werte werden auch als *Hyperparameter* bezeichnet. Die Auswahl der richtigen Hyperparameter erfordert in der Regel viel Geschick und Übung. Neben der händischen Auswahl dieser Hyperparameter existieren jedoch bereits Ansätze unter dem Oberbegriff *Automated machine learning (AutoML)*, um auch diesen Schritt zu automatisieren.

Abhängig von der Architektur des Netzes, der Art erlaubter Kanten und weiterer Eigenschaften lassen sich unterschiedliche Arten von Netzen unterscheiden, die für verschiedene Aufgaben geeignet sind. *Convolutional Neural Networks (CNNs)* [Gu+18] eignen sich insbesondere für

Aufgaben im Bereich der Bilderkennung, *Recurrent Neural Networks (RNNs)* [RHW86] für Aufgaben, bei denen wie beispielsweise in der Sprachverarbeitung zeitliche Abfolgen betrachtet werden müssen, und *Generative Adversarial Networks (GANs)* [Goo+14] zum Erzeugen von künstlichen Datensätzen mit großer Ähnlichkeit zu den Trainingsdaten.

Im Gegensatz zu vielen der bereits vorgestellten Verfahren sind die Vorgänge in Neuronalen Netzen schwieriger nachzuvollziehen und zu erklären. Weiterhin benötigen sie im Normalfall eine große Menge von Trainingsdaten und der Ressourcenbedarf im Trainingsprozess liegt deutlich über dem anderer Methoden. Durch ihre hohe Genauigkeit und beinahe universelle Anwendbarkeit bei geeigneter Auswahl von Netztyp und Architektur sind sie heute jedoch trotzdem in unterschiedlichen Anwendungsdomänen weit verbreitet.

4 Erklärbares ML

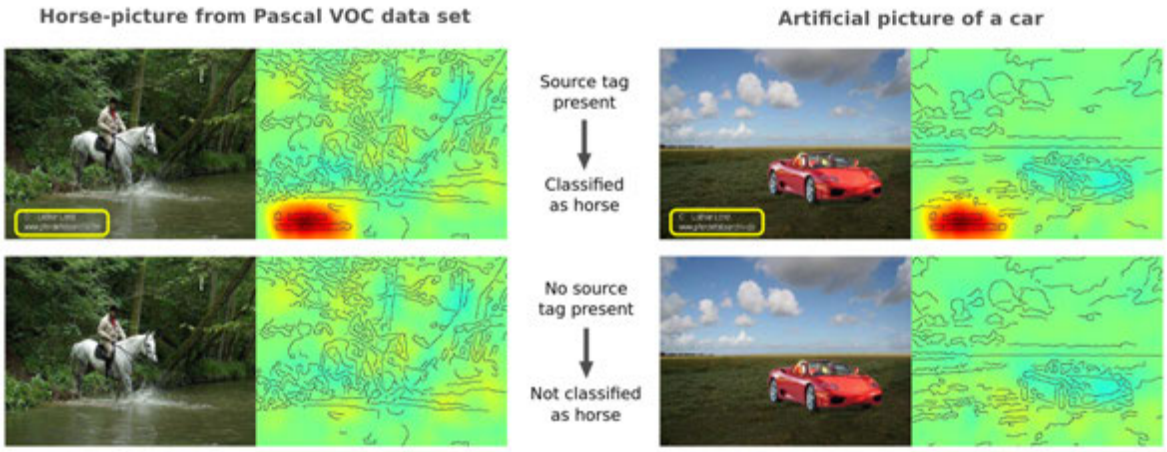

Abbildung 10: Im Rahmen der *PASCAL Visual Object Classes Challenge 2007* wurde ein Modell für die Objektklassifizierung in Bilddaten erstellt. Da viele der enthaltenen Bilder von Pferden den Namen des Fotografen enthielten, lernte das Modell diesen Zusammenhang. Wurde der Namenszug auf das Bild eines Autos gesetzt, so wurde auch dieses fälschlicherweise als Pferd klassifiziert. Durch eine Heatmap-basiertes Erklärungsverfahren wird dieser ungewollte Zusammenhang deutlich. Abbildung entnommen aus [Lap+19].

Erklärbares Maschinelles Lernen (engl. *Interpretable or Explainable ML/AI, XML/XAI*) beschreibt Methoden, die die von einem ML-Verfahren getroffenen Entscheidungen oder erhaltenen Ergebnisse für menschliche Benutzer nachvollziehbar machen [DK17]. Erklärbarkeit ist keine notwendige Eigenschaft, falls der Einsatz von fehlerhaften ML-Modellen keine signifikanten Auswirkungen im Einsatzkontext hat. Gerade beim Einsatz neuer Verfahren in kritischen Kontexten unterstützt die Nachvollziehbarkeit eines Modells jedoch beim Verstehen der Entscheidungen. Dies kann das Vertrauen in die Verfahren stärken oder ihren Einsatz in kritischen Szenarien überhaupt erst ermöglichen. Die Nachvollziehbarkeit kann auch dabei unterstützen, Bias in Modellen aufzudecken und so auch zwischen einfacher Korrelation und Kausalität unterscheiden zu können. Ein Beispiel für eine ungewollte Korrelation, die durch den Einsatz von Erklärverfahren aufgedeckt werden konnte, ist in Abbildung 10 abgebildet. Die Nachvollziehbarkeit kann damit insgesamt zu robusteren Modellen führen, aber auch ein tieferes Verständnis für bisher unbekannte Zusammenhänge fördern [Hol18]. Dieses Problemfeld ist auch im Hinblick auf die

in der DSGVO geforderte verständliche Erklärung der Entscheidungsfindung in automatisierten Entscheidungsprozessen (Artikel 13-15, 22) relevant [Hol+17].

Die Verfahren des erklärbaren maschinellen Lernens lassen sich auf verschiedene Weisen kategorisieren [Mol21]. *Lokale* Erklärungsmodelle bieten Erklärungen dafür, wie ein Modell eine Vorhersage für einen spezifischen Datensatz getroffen hat, *globale* Erklärungsmodelle bieten Erklärungen dafür, wie ein Modell insgesamt Entscheidungen trifft oder zumindest welche Auswirkungen bestimmte Teile eines Modells (beispielsweise einzelne Gewichte) hervorrufen. Es kann weiterhin zwischen *intrinsisch erklärbaren Verfahren* und *Post-hoc-Erklärungen* unterschieden werden. Intrinsisch erklärbare Verfahren sind in ihrer Struktur so einfach, dass sie als menschlich interpretierbar angesehen werden. Beispiele hierfür sind Entscheidungsbäume oder einfache lineare Regressionen. Post-hoc-Erklärungen sind Methoden, die nach der Trainingsphase eines Modells angewendet werden und beispielsweise die Wichtigkeit einzelner Eingangsfeatures bewerten können. Eine weitere Möglichkeit der Differenzierung besteht zwischen *modellspezifischen* und *modellagnostischen* Verfahren – je nachdem, ob Erklärungen nur für bestimmte Arten von Modellen generiert werden können oder ein generischer Ansatz verfolgt wird.

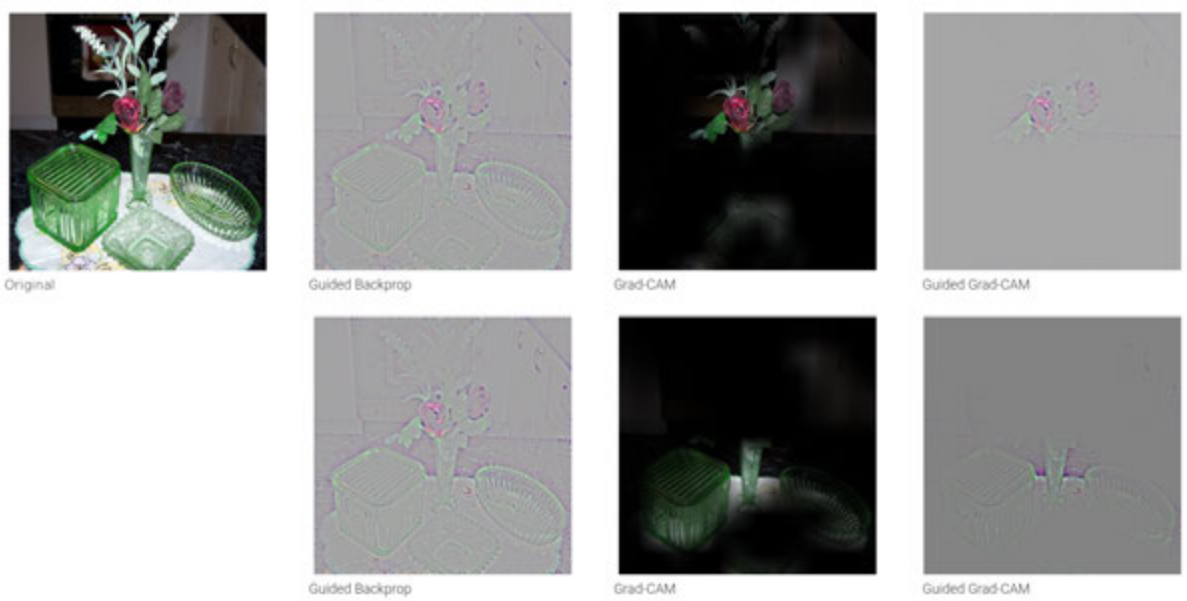

Abbildung 11: Visualisierung des Grad-CAM-Verfahrens für ein LSTM-basiertes *Visual-Question-Answering*-Verfahren. Die obere Bildreihe zeigt die Ausgabe von Grad-CAM für die Frage nach der Farbe der Rosen (Antwort *pink*), die untere die Ausgabe für die Vasenfarbe (Antwort *green*). Eine interaktive Demo, auf der diese Abbildung basiert, ist unter http://gradcam.cloudcv.org/ zu finden.

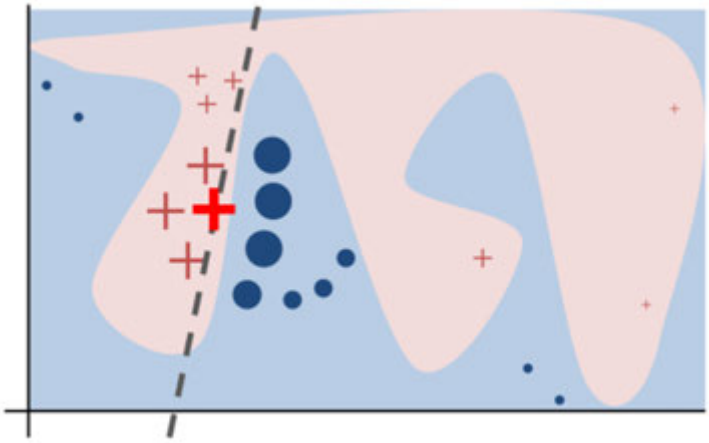

Abbildung 12: Eine Veranschaulichung der Funktionsweise von *LIME* für eine binäre Klassifikation (aus: [RSG16]). Der Entscheidungsraum wird von dem Framework in die beiden Klassen *rot* und *blau* geteilt und basierend auf mehreren Anfragen an das Modell (dargestellt als Markierungen) skizziert. Neben der globalen Skizze wird für einen bestimmten Datenpunkt (hellrote Markierung) außerdem eine *lokale* Erklärung generiert. Diese ist als gestrichelte Linie zu erkennen und erhebt keinen Anspruch an globale Gültigkeit. Abhängig von der Nähe zum Ausgangspunkt werden die Einflüsse für die lokale Erklärung entsprechend gewichtet – hier angedeutet in der Größendarstellung der Markierung.

Ein beliebtes lokales und modellagnostisches Verfahren ist *Local interpretable model-agnostic explanations (LIME)*, das anhand einzelner Modellentscheidungen versucht, eine Skizze des gesamten Entscheidungsraumes zu konstruieren [RSG16]. Für einen einzelnen Datenpunkt kann *LIME* lokale Erklärungsansätze bieten und somit bspw. erklären, welche Merkmale besonders ausschlaggebend für die Entscheidung waren. Die Idee des Verfahrens eingesetzt für ein Klassifikationsproblem besteht darin, zu einem als Black-Box betrachteten Modell, beispielsweise einem trainierten Neuronalen Netz, und einem zu klassifizierenden Datensatz D_k ein interpretierbares Modell, wie etwa einen einfachen Entscheidungsbaum, zu trainieren, das für den speziellen Datensatz D_k Erklärbarkeit liefert. Hierfür werden leicht veränderte Varianten von D_k durch das Netz klassifiziert und die entstandenen Daten als Trainingsdaten für das Training des interpretierbaren Modells genutzt. Abbildung 12 veranschaulicht die Funktionsweise von *LIME* anhand einer binären Klassifikation.

Neben *LIME* existieren verschiedene andere Ansätze, die erklärbares ML ermöglichen. Drei von ihnen werden im Folgenden kurz vorgestellt.

- *Shapley additive explanations (SHAP)* ist ein lokales, modellagnostisches Verfahren, das auf Shaply-Werten – einem Prinzip aus der Spieltheorie – aufbaut, um den Einfluss zu bestimmen, den bestimmte Features des Eingangsdatensatzes auf die Vorhersage des Modells besitzen [LEL19].

- *Gradient-weighted class activation mapping (Grad-CAM)* ist ein lokales, visuelles Erklärungsverfahren für CNN-basierte Modelle, das Heatmaps nutzt, um für die Entscheidung eines Modells relevante Bereiche eines Eingangsbildes zu visualisieren [Sel+17]. Eine beispielhafte Ausgabe des Verfahrens ist in Abbildung 11 zu finden.

- *Layer-Wise Relevance Propagation (LRP)* ist ein lokales Erklärungsverfahren für Neuronale Netze, bei dem die Vorhersage des Modells durch die Schichten des Netzes zurückverfolgt wird und das so letztlich der Einfluss der Eingangsdaten bestimmen kann [Bac+15].

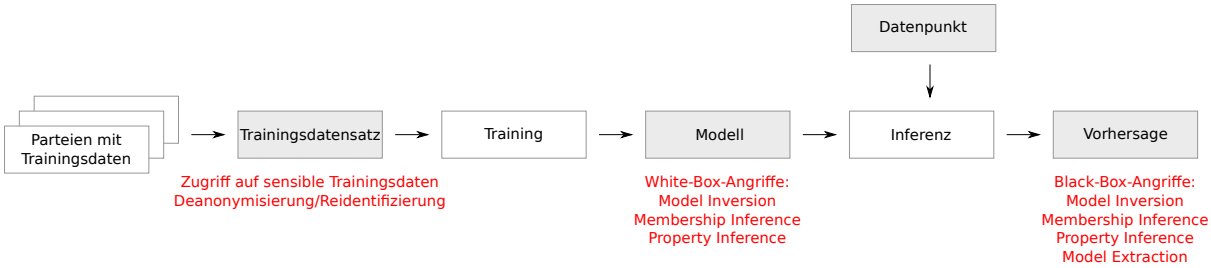

Abbildung 13: Übersicht über Bedrohungen für personenbezogene Daten und Angriffe auf ML-Verfahren.

Eine Übersicht über den aktuellen Stand der Erklärbarkeit von ML-Verfahren und auch beispielhafte Anwendungsfälle aus dem medizinischen Bereich lassen sich in [Kra+21] finden. Im Gegensatz zu diesen technischen Methoden existieren weiterhin Rahmenwerke, die sich auf die Semantik und den Kontext eines Modells konzentrieren [Geb+18; Mit+19]. Dies sind zwar geeignete Mittel, um (zusätzliche) Transparenz zu schaffen, das Zustandekommen von KI-Entscheidungen wird dadurch allerdings nur im Ansatz erklärt.

5 Angriffe auf maschinelle Lernverfahren

In diesem Abschnitt werden Bedrohungen für personenbezogene Daten im Zusammenhang mit maschinellen Lernverfahren sowie ML-spezifische Angriffe dargestellt. Hierbei werden explizit Vertraulichkeitsverletzungen, also beispielsweise die unberechtigte Kenntnisnahme sensibler Gesundheitsdaten, fokussiert. Weitere datenschutzrechtliche Schwächen wie mangelnde Transparenz oder die Gefahr von Profiling sollen hier nicht betrachtet werden. Abbildung 13 bietet einen Überblick über addressierte Bedrohungen, die in den folgenden Abschnitten detailliert dargestellt werden.

Die ersten Bedrohungen beziehen sich auf allgemeine Gefahren beim Umgang mit personenbezogenen Daten, die nicht spezifisch für ML-Verfahren sind. Hierbei steht die potentielle Sensibilität von Trainingsdaten und die Bedrohung durch unberechtigten Zugriff oder unzureichende technische Schutzmaßnahmen im Vordergrund.

Weitere Bedrohungen stellen eigens entwickelte Angriffe auf ML-Modelle dar. Meist liegt dabei folgendes Angriffszenario zugrunde: Ein Angreifer hat Zugriff auf ein bereits trainiertes ML-Modell und versucht, diesem Informationen zu extrahieren, die über die gewöhnliche Anwendung (bspw. Klassifizierung) des Modells hinausgehen. Die beiden Angriffe *Model Inversion* und *Membership Inference* zielen darauf ab, sensible Informationen aus einem Modell zu extrahieren, deren Offenlegung von den Modell-Entwicklern nicht beabsichtigt war. Damit kann die Privatsphäre von den Personen, deren Daten für das Trainieren des Modells verwendet wurden, gefährdet sein. Wenn ein Modell mit privaten Daten trainiert wurde, es aber im Anschluss von anderen Parteien genutzt wird, können Privatsphäreangriffe großen Schaden anrichten.

Dabei wird grundsätzlich zwischen zwei gegensätzlichen Angriffsszenarien unterschieden: *White-Box* und *Black-Box*. Im *Black-Box*-Szenario kann ein Angreifer zwar ein ML-Modell benutzen, um Inferenzen (bspw. Vorhersagen oder Klassifizierungen) durchzuführen. Wie bei einem Orakel kann der Angreifer Datensätze an das Modell schicken, um die Modellausgaben zu erhalten. Ihm steht dabei aber lediglich ein Interface zur Verfügung, hinter dem die Modellinterna, also die Hyperparameter und die antrainierten Parameter, versteckt sind. Im Gegensatz dazu bedeutet *White-Box*-Zugriff volle Einsicht in das ML-Modell: Neben der Möglichkeit, mit dem Modell Inferenzen durchzuführen, hat der Angreifer hiermit Vollzugriff auf das Modell,

inklusive der Einsicht in alle Parameter und Hyperparameter. Als Abstufung zwischen diesen beiden Extremen sind verschiedene *Gray-Box*-Szenarien möglich. Hierzu könnte beispielsweise die Einsicht in manche Hyperparameter oder die Architektur eines Neuronalen Netzes zählen, ohne Zugriff auf andere Parameter.

Die meisten ML-Privatsphäreangriffe sind per Black-Box Zugriff realisierbar. Das bedeutet, dass Angreiferinnen keinen direkten Zugriff auf die Gewichte und Parameter des Modells haben, aber beliebige Anfragen an das Modell stellen können und daraufhin seine Ausgabewerte erfahren. Dies ist bei *Machine Learning as a Service* (MLaaS) oftmals der Fall. MLaaS ist ein Sammelbegriff für (meist kostenpflichtige) Dienstleistungen, die bei der Entwicklung von ML-Modellen unterstützen. Angefangen bei der Installation von Software, über das Preprocessing von Daten bis zur Trainingsphase selbst können so ressourcenintensive Schritte eingespart werden. MLaaS-Provider können also neben Hardware, vorinstallierten Arbeitsumgebungen und Algorithmen ihren Kundinnen auch fertige Modelle zur Verfügung stellen. Neben einigen spezialisierten ML-Plattformen (BigML, Domino, HPE Haven on Demand, Arimo, . . .) bieten mittlerweile auch alle großen Cloud-Computing-Anbieter wie Microsoft Azure, Amazon AWS, IBM Watson und Google Cloud MLaaS an.

5.1 Zugriff auf sensible Trainingsdaten

Ein grundsätzliches Problem beim Einsatz maschineller Lernverfahren in bestimmten Einsatzkontexten besteht in der Sensibilität der verwendeten Trainingsdaten. Werden ML-basierte Systeme beispielsweise in der Medizin eingesetzt, so handelt es sich bei den Daten, die zum Training verwendet werden, häufig um sensible Gesundheitsdaten. Die einfachste Möglichkeit für den Einsatz eines ML-basierten Verfahrens ist die Sammlung der Daten (ggf. aus mehreren Quellen) an einem zentralen Ort und das anschließende Training auf diesen Daten. Werden keine geeigneten Schutzmaßnahmen getroffen, so können der Betreiber des Verfahrens oder im schlimmsten Fall sogar Externe direkten Zugriff auf die sensiblen Trainingsdaten erhalten. Wenn die Daten einen Personenbezug ermöglichen, so kann dies abhängig von der Art der Daten zu direkten negativen Auswirkungen auf Betroffene führen.

5.2 Deanonymisierung/Reidentifizierung

Eine häufig verwendete Maßnahme zum Schutz von personenbezogenen Daten vor dem Training besteht in der Anonymisierung oder Pseudonymisierung der Daten (siehe Abschnitt 6.1). In vielen Fällen ist es trotz dieser Maßnahmen möglich, einen Personenbezug wiederherzustellen. Ein Beispiel hierfür ist ein von Netflix veröffentlichter, vermeintlich anonymisierter Datensatz zu Filmbewertungen, der es durch Kombination mit öffentlich verfügbaren Daten der *Internet Movie Database* ermöglichte, einzelne Benutzer zu identifizieren und Aussagen über ihre politische Einstellung und sexuelle Orientierung zu treffen [NS08]. Im Einzelfall müssen existierende Risiken und verwendete Maßnahmen genauestens bewertet und ausgewogen werden.

5.3 Model Inversion

Abbildung 14: Ein Beispiel des *Model Inversion* Angriffs aus [FJR15] mit einer Rekonstruktion (links) des Originalbilds (rechts) aus den Trainingsdaten.

Durch *Model Inversion* können Angreifer per Black-Box-Zugriff Teile der Trainingsdaten rekonstruieren, ohne zuvor tiefgreifendes Wissen über die verwendeten Daten zu haben. Ein bekanntes Beispiel ist in Abbildung 14 zu sehen, bei dem Fredrikson et al. ausgehend von einem zufällig generierten Bild ein Portraitfoto mit erkennbaren Merkmalen aus den Trainingsdaten eines Gesichtserkennungsmodells rekonstruieren konnten [FJR15]. In bestimmten Szenarien genügt es somit, wiederholt gezielte Anfragen an ein Modell zu stellen, um erfolgreich Informationen aus den Trainingsdaten zu rekonstruieren. Diese Art des Angriffs wird seit der Erstveröffentlichung weiter erforscht und es wurden bereits verschiedene Varianten und Weiterentwicklungen veröffentlicht (z.B. [Zha+20]). Allerdings scheinen erfolgreiche *Model-Inversion*-Angriffe nur in bestimmten Szenarien möglich zu sein: So zeigen etwa Shokri et al., dass der Angriff lediglich den Durchschnitt der Trainingsdaten einer Datenklasse aufdecken kann, was in vielen Anwendungsfällen zu unbrauchbaren Ergebnissen führt (siehe Abbildung 15) [Sho+17].

Abbildung 15: Die Ergebnisse von Shokri et al. [Sho+17], die den *Model Inversion*-Angriff [FJR15] auf einem Bilderkennungsmodell für den Datensatz CIFAR-10 ausgeführt haben. Hier sind Rückschlüsse auf die Trainingsdaten kaum möglich: Die Darstellungen der oberen Reihe korrespondieren mit den jeweiligen durchschnittlichen Trainingsdaten der Klassen Flugzeug, Auto, Vogel, Katze und Reh (v.l.n.r.). Die untere Reihe zeigt die Ergebnisse für die Klassen Hund, Frosch, Pferd, Schiff, LKW.

5.4 Membership Inference

Das Ziel eines *Membership-Inference*-Angriffs [Sho+17] ist es, für einzelne Datenpunkte zu entscheiden, ob sie Teil des Trainingsdatensatzes eines ML-Modells waren. Auch dieser Angriff kann per Black-Box-Zugriff realisiert werden: Zugrunde liegt hierbei die Idee, dass sich die Ausgaben des Modells für arbiträre Daten etwas von Ausgaben für jene Daten unterscheiden, mit denen das Modell trainiert wurde. Da sich das Modell während des Trainings schrittweise an die Trainingsdaten angepasst hat, sind Vorhersagen für Trainingsdaten in der Regel exakter bzw. eindeutiger. Eine Überanpassung (*Overfitting*, siehe Abschnitt 2.4) begünstigt den Angriff daher zusätzlich. Eine zunächst wirksame Methode gegen Membership-Inference-Angriffe ist es, die Ausgabe von ML-Modellen zu reduzieren, sodass bspw. in einer Klassifizerungsaufgabe statt Wahrscheinlichkeitswerten für Klassenzugehörigkeiten nur der Name der wahrscheinlichsten Klasse ausgegeben wird („label-only"). Jedoch sind Membership-Inference-Angriffe gut erforscht und es wurden zahlreiche Varianten entwickelt, die bspw. auch für label-only ML-Algorithmen funktionieren [Cho+21].

In der Literatur werden die beiden Fälle *Zugehörigkeit* und *Nicht-Zugehörigkeit* eines Datums zu den Trainingsdaten in der Regel scharf voneinander abgegrenzt: Angreifer evaluieren ihren Angriff für Daten, die exakt in der vorliegenden Form Teil der Trainingsdaten waren, bzw. für Daten, die sich maßgeblich von Trainingsdaten unterscheiden. Membership Inference wäre aber auch Daten denkbar, die Trainingsdaten *ähnlich* sind – bspw. für Fotos aus einer leicht veränderten Perspektive oder tabellarische Daten mit geringer Unschärfe. Angreifer könnten dann mehrere Anfragen an das Modell stellen, jeweils mit verschiedenen Varianten der Informationen, die über ein potenzielles Trainingsdatum vorliegen. Eine systematische Auswertung der Modell-Rückgabewerte können dann ebenfalls Rückschlüsse auf Trainingsdaten zulassen.

Das Anwendungsgebiet eines ML-Algorithmus ist hierbei ausschlaggebend dafür, inwieweit der Angriff zu Privatsphäreverletzungen führen kann. Handelt es sich um einen Algorithmus, die beispielsweise eine Wahrscheinlichkeit dafür errechnet, schwere Straftaten zu begehen oder an einer schweren Krankheit zu erkranken, kann die bloße Tatsache, dass die Daten einer Person im Trainingsdatensatz enthalten sind, bereits privatsphäreverletzend sein.

5.5 Property Inference

Ähnlich wie Membership Inference, zielen auch *Property Inference* Angriffe auf die einem Modell zugrunde liegenden Trainingsdaten ab [Ate+15]. Die meisten Property Inference Angriffe haben das Ziel, globale Eigenschaften der Trainingsdaten anhand des Modellverhaltens (im Black-Box Szenario) oder anhand der Modellparameter (im White-Box Szenario) zu extrahieren. Da diese Eigenschaften in der Regel nicht in direktem Zusammenhang zu der vom Modell gelernten Aufgabe stehen, können sie mehr über das Modell bzw. dessen Trainingsdaten verraten, als vom Modellbesitzer beabsichtigt.

Beispielsweise können per Property Inference die Geschlechterverteilung und andere statistische Eigenschaften in tabellarischen Trainingsdaten, das Vorhandensein von Rauschen in Bildern, oder der Anteil von Bildern mit bestimmten Merkmalen wie Alter oder Geschlecht in den Trainingsdaten aus Modellen extrahiert werden [Gan+18]. Für kollaborative Lernszenarien (z.B. Federated Learning, siehe Abschnitt 6.3) sind Property Inference Angriffe von besonderer Bedeutung, da hier die Privatsphäre einzelner Teilnehmer explizit geschützt werden soll und das Extrahieren von Trainingsdateneigenschaften diese verletzt [Mel+19].

5.6 Model Extraction

Model Extraction ist ausschließlich für das Black-Box-Szenario konzipiert. Ziel des Angriffs ist hierbei die Übertragung des Zielmodell-Verhaltens auf ein eigenes Modell, um die aufwendige Trainingsphase für das eigene Modell zu umgehen [Pap+17]. Dies ist insbesondere für den MLaaS-Kontext (siehe Abschnitt 5) relevant, da hier oft für den Zugriff auf ein Modell bezahlt werden muss. Durch das Anfertigen eines gleichwertigen Modells per *Model Extraction* kann dies unter Umständen umgangen werden. Somit zielt dieser Angriff eher auf Betriebsgeheimnisse ab als auf sensible Informationen, die aus datenschutzrechtlichen Gründen geschützt werden müssen.

6 Privacy-Preserving Machine Learning

Privacy-preserving Machine Learning (ppML, dt. etwa „Privatsphäre-wahrendes Maschinelles Lernen“) ist ein Überbegriff für verschiedene Techniken, die dazu beitragen können, Maschinelles Lernen privatsphärefreundlicher zu gestalten. Je nach Bedarf können diese Techniken verschiedene Schutzziele sowohl in der Trainingsphase als auch in der Inferenzphase verfolgen. Beispielsweise kann der Einsatz von *Federated Learning* die zum Training verwendeten Daten vor dem Betreiber eines Systems schützen, wohingegen *homomorphe Verschlüsselung* in der Inferenzphase die zu bewertenden Daten vor dem Betreiber schützen kann. Abbildung 16 bietet einen Überblick darüber, an welchen Stellen welche Mechanismen für verschiedene Schutzziele eingesetzt werden können. Die einzelnen Verfahren werden in den folgenden Abschnitten näher vorgestellt. Eine Übersicht über die vorgestellten Verfahren befindet sich in Tabelle 1.

6.1 Pseudonymisierung und Anonymisierung

Die *Pseudonymisierung* und *Anonymisierung* von Daten vor ihrer Nutzung in der Trainingsphase eines ML-Verfahrens zum Schutz der Betroffenen stellt eine weit verbreitete Maßnahme dar. Pseudonymisierung beschreibt dabei die Ersetzung direkt identifizierender Merkmale, beispielsweise eines Namens, durch Pseudonyme, häufig ausreichend lange zufällige Zeichenketten. Die Zuordnung des identifizierenden Merkmals zu dem Pseudonym kann in Tabellenform gespeichert werden und eine Re-Identifizierung der Betroffenen erlauben. Durch die Verwendung von Pseudonymen kann die Verkettbarkeit verschiedener Datensätze ohne eine Offenlegung direkt identifizierender Merkmale ermöglicht werden.

Anonymisierung beschreibt die Veränderung der Daten derart, dass sie nicht mehr oder nur mit unverhältnismäßig hohem Aufwand einer Person zugeordnet werden können. Hierbei

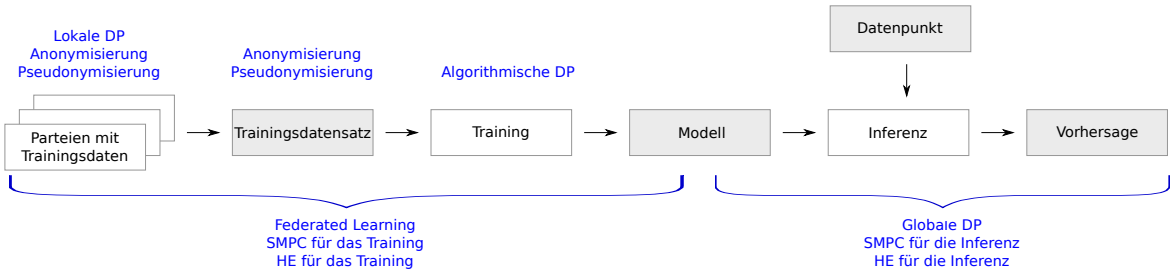

Abbildung 16: Verfahren des *Privacy-Preserving Machine Learning* und wo sie zum Einsatz kommen.

Technik	Phase der Anwendung	Funktion	Schwächen
Pseudonymisierung und Anonymisierung	Vor dem Training	Verhindert Rückschlüsse auf Betroffene	Gefahr von Reidentifizierung
Differential Privacy	Training und/oder Inferenz	Minimiert Einfluss einzelner Datenpunkte	Mindert Transparenz, nicht für alle Datentypen geeignet
Federated Learning	Verteiltes Training	Gemeinsames Training ohne Rohdatenaustausch	Viele Angriffe weiterhin möglich
Secure Multiparty Computation	Verteiltes Training und/oder Inferenz	Gemeinsame Funktionsberechnung ohne Rohdatenaustausch	Häufig rechenintensiv, hohe Einstiegshürden
Homomorphe Verschlüsselung	Training und/oder Inferenz	Berechnungen auf verschlüsselten Daten	Häufig sehr rechenintensiv (FHE) oder Ersetzung bestimmter Funktionen notwendig
Trusted Execution Environments	Training und/oder Inferenz	Hardwareunterstütze Absicherung gegen lokale Angreifer	Nur mit geeigneter Hardware möglich, Angreifbarkeit über Seitenkanäle

Tabelle 1: Übersicht über die vorgetellten ppML-Verfahren inklusive ihrer Anwendungsphase, Funktion und Schwächen.

können verschiedene Methoden zum Einsatz kommen, die von der Art der Daten und dem Einsatzzweck abhängen, u. a.:

- Daten können *unterdrückt*, d. h. vollständig entfernt, werden.
- Durch das *Verrauschen*, beispielsweise durch das Hinzufügen normalverteilter Zufallswerte, lassen sich allgemeine Verteilungen erhalten, aber Daten einzelner Personen unschärfer machen.
- Die *Generalisierung* von Daten bezeichnet ihre Ersetzung durch Vergröberungen, beispielsweise die Angabe eines Geburtsjahres anstelle des Geburtsdatums.
- Die *Aggregation* beschreibt die Nutzung von kombinierten Daten, wie beispielsweise Durchschnittswerten, anstelle von Einzeldatensätzen.

Um die Güte einer Anonymisierung bewerten zu können, können verschiedene Metriken wie k-Anonymität oder l-Diversität herangezogen werden [Swe02; Mac+06]. Es existieren verschiedene Verfahren, wie beispielsweise Mondrian [DLR06], für die Anonymisierung von Datensätzen, die diesen Metriken genügen. Ebenso existieren Ansätze dafür, den Anonymisierungsprozess für mehrere Parteien nicht zentral, sondern verteilt zu gestalten, sodass keine Partei in den Besitz eines vollständigen nicht-anonymisierten Datensatzes gelangen muss [JX09].

Bei der Anonymisierung von Daten muss zwischen der Nützlichkeit der Daten und dem Risiko einer Re-Identifizierung von Betroffenen abgewogen werden (der sogenannte *Risk-Utility*

Trade-off). Zu stark anonymisierte Daten schließen zwar eine Re-Identifkation aus, können aber auch zu wenig Information enthalten, um noch für das Trainieren eines ML-Modells nutzbar zu sein.

Der Vorteil von Pseudonymisierung und Anonymisierung ist häufig ihre Einfachheit, insbesondere etwa wenn beteiligte Parteien nicht direkt miteinander kommunizieren können, wie es für andere Techniken notwendig ist. Jedoch lässt sich in vielen Fällen eine Re-Identifizierung nicht vollständig ausschließen (siehe auch Abschnitt 5.2), sodass der Einsatz entsprechender Maßnahmen sorgfältig abgewägt werden muss.

6.2 Differential Privacy

Differential Privacy (DP) ist eine von Dwork u. a. entwickelte Metrik [Dwo+06], die sich in den vergangenen Jahren zum „de-facto Standard für privatsphärefreundliche Datenanalyse“ entwickelt hat [Pih+18]. Das Ziel von DP ist es, die Genauigkeit der Datenanalyse zu maximieren, während die Identifizierbarkeit einzelner Individuen in den zugrunde liegenden Daten minimiert wird. Dabei folgt DP der Idee, dass Außenstehende bei einer Auswertung auf einem *differentially private* Datensatz nicht unterscheiden können, ob sich die Daten eines bestimmtes Individuums im Datensatz befunden haben oder nicht. Somit können ML-Algorithmen bspw. resistent gegen Privatsphäreangriffe werden, die wie Membership Inference auf einzelne Individuen in Trainingsdatensätzen abzielen (siehe Abschnitt 5.4).

DP kann dabei an verschiedenen Punkten der Datenverarbeitung ansetzen: *Lokale DP* wird auf den Eingabe-Datensatz angewendet, während *globale DP* die Ausgabe eines Algorithmus verändert und *algorithmische DP* Zwischenergebnisse in iterativen Algorithmen betrifft.

Ein Beispiel für algorithmische DP im ML-Kontext haben Abadi u. a. vorgestellt. Ihr Ansatz fügt während des Trainings von neuronalen Netzen in der Lernfunktion (hier: *stochastic gradient descent*) in jeder Iteration gezielt Rauschen hinzu, um Differential Privacy zu erreichen [Aba+16].

Lokale DP kann insbesondere in verteilten ML-Szenarien wie Federated Learning genutzt werden. Dies ermöglicht es den teilnehmenden Datenspendern, ihre Daten entsprechend lokal mit Privatsphäregarantien zu verändern, bevor die eigenen Daten an eine zentrale Stelle, die etwa ein ML-Modell trainiert, weitergeleitet werden [Pih+18]

Konkret bezieht sich die Definition von DP auf *benachbarte* Datensätze D und D', also Datensätzen, die sich in höchstens einem Datum unterscheiden. Die Ausgaben eines DP-Algorithmus dürfen sich bei der Eingabe der beiden Datensätze D und D' nur leicht unterscheiden, wobei der erlaubte Spielraum vom sogenannten Privatsphärebudget ε und von einem weiteren Parameter δ festgelegt wird.

Formal gilt ein randomisierter Algorithmus A als (ε, δ)-*differentially private*, wenn für alle möglichen Untermengen der Ausgabewerte $Y \subseteq Range(A)$ gilt:

$$\Pr\left[A(D) \in Y\right] \leq e^{\varepsilon} \Pr\left[A(D') \in Y\right] + \delta$$

Die Wahrscheinlichkeiten beziehen sich dabei auf die zufälligen Entscheidungen von A. Während $\delta = 0$ als *pure* DP bezeichnet wird, erlaubt $\delta > 0$ dem Algorithmus, mit geringer Wahrscheinlichkeit nicht *differentially private* zu sein. Je kleiner ε gewählt ist, desto weniger Aussagen lassen sich anhand der Ausgaben von A über einzelne Datenpunkte treffen. $\varepsilon = 0$ (in Kombination mit $\delta = 0$) würde den Betroffenen somit zwar maximale Privatsphäre bieten, aber gleichzeitig den Algorithmus A unbrauchbar machen, da unterschiedliche Eingabedaten keine Auswirkungen mehr auf seine Ausgaben hätten.

DP bringt drei wichtige Eigenschaften mit sich: Einerseits ist DP *kompositionsfähig* (engl. *composable*), sodass die Komposition mehrerer DP-Mechanismen weiterhin *differentially private* ist. Andererseits ist DP robust gegenüber Hintergrundinformationen: Die Privatsphäregarantie von DP ist unabhängig von möglichem Hintergrundwissen eines Angreifers, sodass ein Kombinieren der Algorithmenausgaben mit anderen Datenquellen keinen Erkenntnisgewinn aus Angreifersicht ermöglicht. Außerdem bietet DP Gruppenprivatsphäre (engl. *group privacy*), sodass korrelierte Eingaben (bspw. mehrere Datensätze zu einer Person) die Privatsphäregarantien nicht übermäßig abschwächen [Aba+16; Dwo+06].

Allerdings kann die Anwendung von DP in ML-Algorithmen dahingehend problematisch sein, dass die ohnehin schwer erklärbaren Prozesse des ML durch DP an zusätzlicher Komplexität gewinnen und für Menschen schwerer nachvollziehbar werden. Außerdem sind die Anwendungsmöglichkeiten mancher DP-Techniken auf bestimmte Datentypen beschränkt. Während die Reihenfolge tabellarischer Daten bspw. problemlos neu gemischt werden kann, ist dies für die Pixelreihenfolge auf Bildern nicht ohne Weiteres möglich. Obwohl bei Bilddaten geringfügiges Verrauschen, eine gängige Praxis zum Erreichen von DP, technisch ohne Weiteres möglich ist, sind die Auswirkungen von verrauschten Bilddaten für ML-Algorithmen mitunter unklar: Einerseits gibt es Angriffe gegen ML-Modelle mit sogenanntem *adversarial noise* (etwa: „bösartiges Rauschen“) in Trainingsdaten, wodurch ein Angreifer den Trainingsprozess bzw. das Modellverhalten zu seinen Gunsten manipulieren kann [GSS14]. Andererseits wurde gezeigt, dass das Hinzufügen von Rauschen in Trainingsdaten als Regularisierungsmethode eingesetzt werden kann, wodurch ein Modell weniger zur Überanpassung (siehe Abschnitt 2.4) neigt [You+19]. Vor einem großflächigen Einsatz von DP, etwa im Bereich der medizinischen Bildanalyse, muss daher noch weitere Forschung betrieben werden [Kai+20].

6.3 Federated Learning

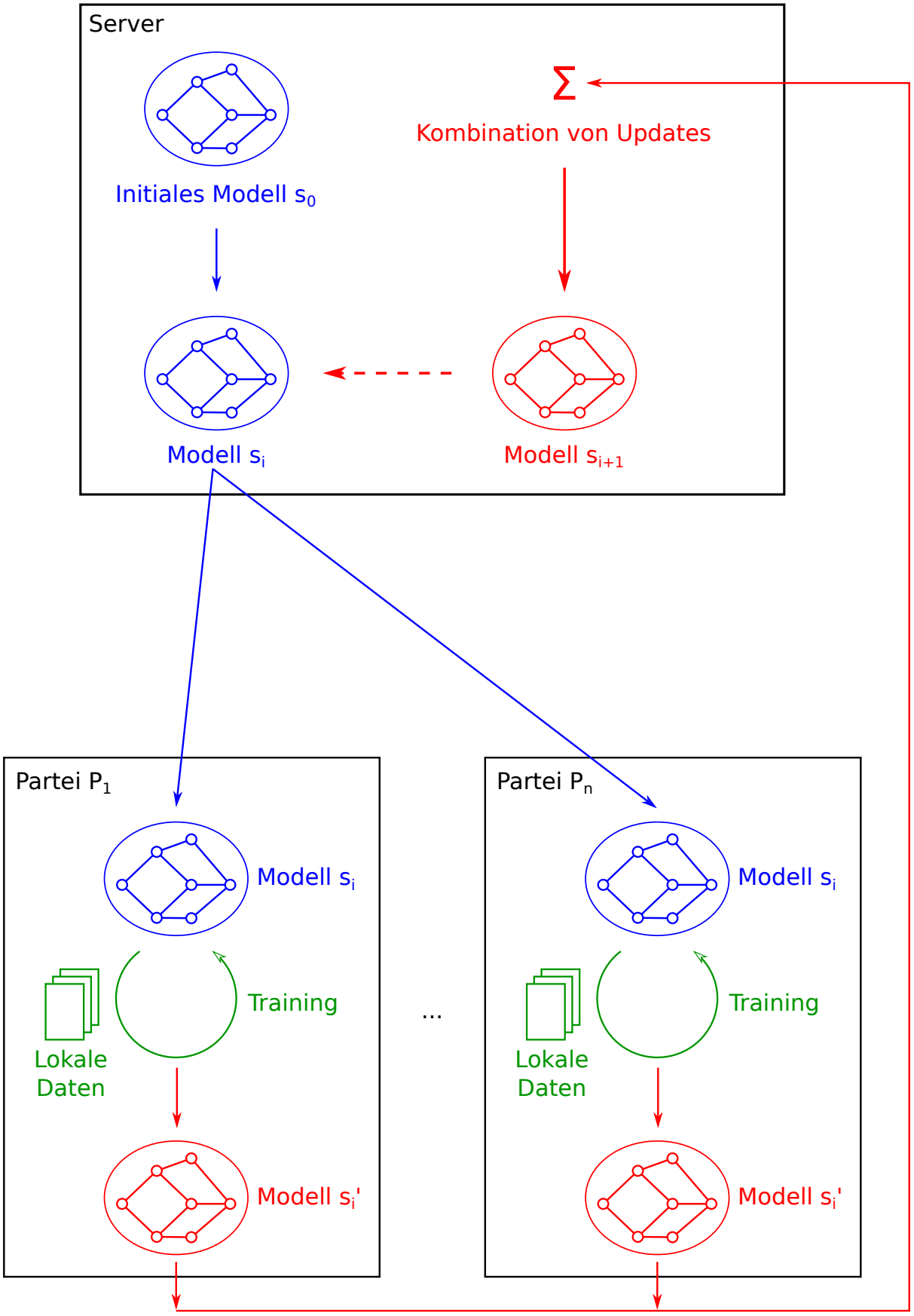

Abbildung 17: Eine vereinfachte Darstellung des Federated-Learning-Prozesses mit einem zentralen Server und n Parteien.

Federated Learning (FL) folgt dem *Edge-Computing*-Paradigma, bei dem Daten dezentral am Rand eines Netzwerkes, oft direkt innerhalb der datenerhebenden Entität, verarbeitet werden. Edge Computing erfreut sich aus verschiedenen Gründen immer größerer Beliebtheit: Einerseits sind die Speicher- und Rechenkapazitäten in Endgeräten (z.B. Smartphones) rapide gewachsen, sodass diese zusätzliche Berechnungen vornehmen können, ohne die Qualität anderer Services einschränken zu müssen. Andererseits sorgt der Trend zu modernen IoT-Netzwerken für immer mehr potenzielle Datenquellen, mit deren Daten Erkenntnisse verschiedenster Art generiert werden können.

Durch FL können ML-Modelle mit Daten, die dezentral über mehrere Parteien verteilt sind, trainiert werden. Dabei findet das Trainieren des Modells auf den privaten Daten direkt auf dem Gerät des Datenbesitzenden statt, sodass die Daten weder mit anderen Teilnehmenden, noch mit der zentralen, orchestrierenden Einheit geteilt werden müssen. Es werden lediglich Modellupdates zwischen den Teilnehmenden und einem zentralen Server ausgetauscht.

Das Ziel des Trainingsprozesses bleibt dabei das gleiche wie bei anderen ML-Verfahren, also beispielsweise die Minimierung fehlerhafter Vorhersagen eines neuronalen Netzes. Die Besonderheit von FL ist, dass zwar ein globales Modell trainiert wird, die Optimierungen (in Form von Modellupdates) jedoch lokal berechnet und anschließend vom zentralen Server kombiniert werden.

Die Trainingsphase für FL funktioniert wie folgt: Der Server initialisiert zunächst ein ML-Modell s_0. Der Server wählt nun für jede Trainingsrunde i Parteien aus, die in dieser Runde am Training des Modells beteiligt sind. Diese Parteien laden das aktuelle Modell s_i vom Server und trainieren s_i mit lokalen Daten weiter, wodurch das aktualisierte Modell s_i' entsteht. s_i' wird zurück an den Server geschickt, auf dem das Update s_i' mit den Updates aller Parteien der aktuellen Trainingsrunde zu s_{i+1} kombiniert wird; beispielsweise kann der Durchschnitt aller Updates berechnet werden. In der nächsten Runde wird nun das weiterentwickelte Modell s_{i+1} trainiert. Die Anzahl der Trainingsrunden kann variieren und wird vom zentralen Server festgelegt. Das Training kann (je nach Einsatzgebiet) effizienter sein, wenn die am Training beteiligten Parteien von Runde zu Runde variieren [Li+20; McM+17]. Die Trainingsphase wird in Abbildung 17 dargestellt.

Ein bekanntes Anwendungsbeispiel für FL ist *Gboard*, die Touch-Tastatur von Google [Har+18]. Diese bietet unter anderem die Vorhersage der nächsten Wörter, basierend auf dem bereits eingegebenen Text. An der Modellentwicklung sind hunderte Millionen Android-Geräte beteiligt, die die Modellupdates jeweils nachts während des Akkuladevorgangs berechnen.

Nachfolgend werden einige der größten Herausforderungen beim Design von FL-Algorithmen dargelegt [Li+20]:

Heterogenität: Die Beteiligung einer Vielzahl von Geräten (bspw. im IoT/Sensor-Kontext) kann einerseits dazu führen, dass die für das Training zur Verfügung stehende Rechenkapazität stark variiert und dass der Gesamtprozess des Trainings insgesamt unvorhersehbarer wird. Andererseits können auch die Daten selbst sehr heterogen sein. Während ein gewisses Maß an Diversität zu einer besseren Generalisierbarkeit des ML-Modells führt, muss gleichzeitig sichergestellt werden, dass die einfließenden Daten für das Training geeignet sind. Da nur die Teilnehmenden selbst Einblick in ihre jeweiligen Daten haben, muss diese Entscheidung grundsätzlich auch auf den Geräten getroffen werden.

Fehlertoleranz: Insbesondere der Ausfall eines beteiligten Gerätes, das während einer Trainingsrunde die Verbindung zum Netzwerk verliert, ist in vielen Szenarien des FL keine Seltenheit. Ein anschauliches Beispiel hierfür sind Smartphones, die keinen durchgehenden Empfang haben, oder durch ihre Benutzer zwischenzeitlich manuell vom Netzwerk getrennt

werden. Diese Ausfallrate kann beispielsweise durch Redundanz kompensiert werden, indem deutlich mehr Geräte am Training beteiligt werden als nötig.

Datenmengen: Das Kommunizieren von Modellupdates kann durchaus mehr Zeit als die Berechnungen selbst beanspruchen. Insbesondere in Szenarien, in denen beispielsweise Millionen von Smartphones am Training beteiligt sind, ist der Einsatz von effizienten Methoden zur Kommunikation unabdingbar. Dabei kann sowohl die Rundenanzahl, als auch die jeweilige Nachrichtengröße (die Datenmenge, die benötigt wird, um ein Modellupdate zu kommunizieren) optimiert werden.

Privatsphäre. Auch wenn die Trainingsdaten der Teilnehmenden nicht aggregiert oder weitergegeben werden, können unter Umständen sensible Informationen aus individuellen Modellupdates extrahiert werden; ähnlich zu den Angriffen in Abschnitt 5 [HAP17]. Im Kontext von FL kann zwischen den beiden Definitionen der *lokalen* und der *globalen Privatsphäre* unterschieden werden [Li+20]. Während letztere Definition dem Standardmodell entspricht, das einen vertrauenswürdigen Zentralserver voraussetzt, bleiben die Modellupdates bei *lokalen* Privatsphäregarantien auch vor dem zentralen Server geheim. Dies kann beispielsweise durch gezieltes Verrauschen der Modellupdates mithilfe von *Differential Privacy* (siehe Abschnitt 6.2) oder durch *Secure Multiparty Computation* realisiert werden [Bon+17; Li+20].

Ohne jene Maßnahmen werden Angriffe wie Membership Inference, Property Inference, Model Inversion und weitere grundsätzlich weiterhin möglich [LYY20]. Deshalb sollte FL in kritischen Anwendungen mit anderen ppML-Methoden kombiniert werden [CP21].

6.4 Secure Multiparty Computation

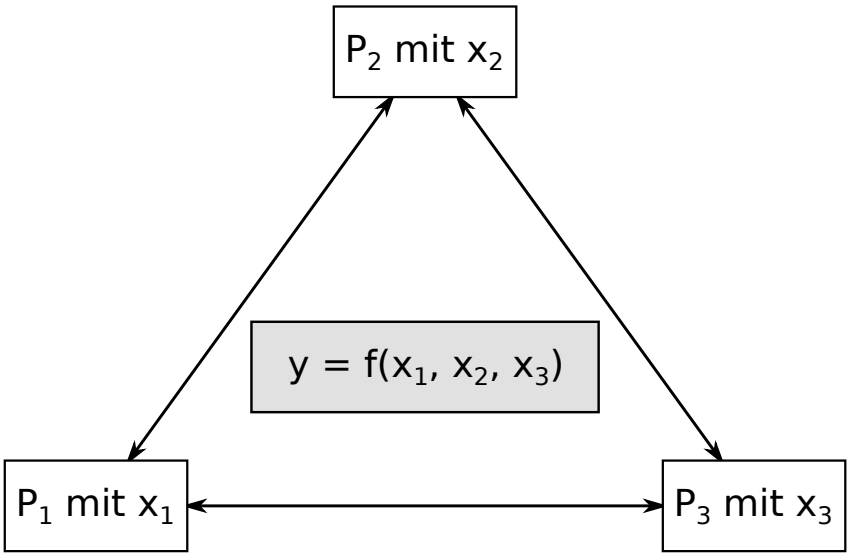

Abbildung 18: Secure Multiparty Computation, hier dargestellt mit drei Parteien, erlaubt die Berechnung einer Funktion f auf den geheimen Eingaben x_i, ohne dass die Parteien die Eingaben anderer Parteien erfahren.

Besonders in verteilten Szenarien kann *Secure Multiparty Computation* (SMPC) eine der Grundlagen für privatsphärefreundliches ML bilden. SMPC erlaubt es mehreren Parteien, eine Funktion gemeinsam zu berechnen, während ihre jeweiligen Eingabedaten geheim bleiben. Nach der Berechnung kennt jede Partei nur ihre eigenen Eingabedaten und den Ausgabewert (bzw. die Ausgabewerte) der Funktion [CD+15].

Formal wollen n Parteien mit ihren jeweiligen Datensätzen $x_1, \ldots, x_n$ die Ausgabe y einer Funktion f berechnen. Im Zuge der Berechnung erlangt keine Partei P_i Zugriff über Informationen, die über ihre eigene Eingabe x_i und das Ergebnis $y = f(x_1, \ldots, x_n)$ hinausgehen. Dieses Prinzip wird in 18 dargestellt.

Um dieses Ziel zu realisieren, kommen in SMPC-Protokollen üblicherweise verschiedene Techniken zum Einsatz. Ein Beispiel für *Secure Two-Party Computation* (2PC) sind sogenannte *Garbled Circuits*, die insbesondere in Szenarien eingesetzt werden können, in denen die Trainingsdaten auf zwei verschiedene Server verteilt sind [MZ17]. So könnten etwa zwei Krankenhäuser ein Modell mit Patientendaten trainieren, ohne dabei die Daten selbst an Dritte weiterzugeben. Im Kontext von Federated Learning kann ein aus dem SMPC-Bereich stammendes *Secure-Aggregation*-Protokoll dabei helfen, die Modellupdates einzelner Geräte zu kombinieren, ohne dabei Einsicht in die individuellen Beiträge zu gewähren (vgl. *lokale Privatsphäre* in Abschnitt 6.3) [Bon+17]. Dies kann es erleichtern, anonyme Nutzerdaten (bspw. von Smartphonenutzenden) privatsphärefreundlich für das Training von ML-Algorithmen zu nutzen.

Neben der Trainingsphase kann auch die Inferenzphase in verteilten Szenarien durch SMPC geschützt werden. So erlaubt bspw. das Framework *MiniONN* [Liu+17] das Überführen beliebiger Neuronaler Netze in SMPC-geschützte Netze für eine verteilte Inferenzphase.

Ein wichtiges Auswahlkriterium für ein geeignetes SMPC-Framework ist das Angreifermodell, vor dem geschützt werden soll. Grundlegend wird zwischen *honest-but-curious* (auch: *semi-honest* oder passiv) und *malicious* (aktiv) unterschieden: Während sich im *honest-but-curious*-Modell alle Teilnehmenden so verhalten, wie es das Protokoll vorschreibt und lediglich die Informationen nutzen, die ihnen preisgegeben werden, können *malicious* Angreifende auch vom Protokoll abweichen und so den Ablauf manipulieren. Die meisten Frameworks (bspw. *MOTION* [Bra+20]) schützen gegen das etwas schwächere *honest-but-curious* Angreifermodell. Ein Schutz gegen aktive Angreifer ist in der Regel aufwendiger, wird aber bspw. im *MP-SPDZ* Framework ermöglicht [Kel20].

Beim Einsatz von kryptografischen Techniken muss stets ein besonderes Augenmerk auf die Recheneffizienz der eingesetzten Algorithmen gelegt werden, insbesondere bei ohnehin anspruchsvollen Operationen wie dem Training eines neuronalen Netzwerks. Eine achtlose Kombination von ML und SMPC kann ansonsten unverhältnismäßig hohe Rechenaufwände nach sich ziehen. Die Praxistauglichkeit von SMPC im ML-Kontext ist einerseits durch die Aufwändigkeit der eingesetzten Rechenoperationen eingeschränkt, andererseits ist auch die Einstiegshürde für Nicht-Kryptografen in bestehende Implementierungen und Frameworks recht hoch [CP21]. Einem großflächigen Einsatz von SMPC stehen diese beiden Aspekte derzeit noch im Weg.

6.5 Homomorphe Verschlüsselung

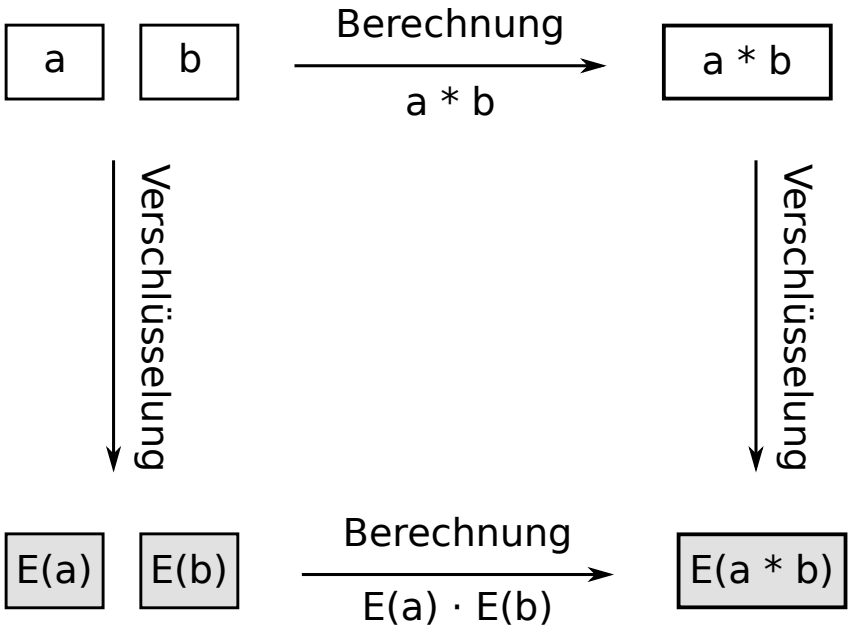

Abbildung 19: Das Prinzip homomorpher Verschlüsselung dargestellt anhand der Multiplikation zweier Zahlen a und b. $E(a)$ beschreibt den Schlüsseltext, der entsteht, wenn a verschlüsselt wird.

In der Mathematik wird ein *Homomorphismus* als eine Abbildung verstanden, die die Elemente einer Menge auf eine andere Menge abbildet und dabei deren mathematische Struktur erhält. Daten, die per Homomorpher Verschlüsselung (engl. *Homomorphic Encryption*, HE) verschlüsselt wurden, behalten ihre Struktur insofern, als dass auf ihrer verschlüsselten Form Berechnungen durchgeführt werden können, die Operationen auf den unverschlüsselten Daten entsprechen. Beispielsweise lässt sich so das verschlüsselte Produkt $E(a*b)$ zweier verschlüsselter Zahlen $E(a)$ und $E(b)$ als $E(a) \cdot E(b)$ berechnen, ohne die Daten entschlüsseln zu müssen. Dieser Zusammenhang wird in Abbildung 19 dargestellt. HE wurde erstmals 1978 von Rivest et. al im Zusammenhang mit privatsphärefreundlichen Datenanalysen vorgeschlagen [RAD+78].

Im Kontext von ML kann HE für verschiedene Verfahren genutzt werden. Dabei können die jeweiligen Berechnungsschritte während der Trainings- und/oder Inferenzphase durch entsprechende HE-Operationen auf verschlüsselten Daten ersetzt werden. So haben bspw. Graepel u. a. mit *ML Confidential* für verschiedene Klassifizierungsverfahren vorgestellt, wie sowohl während der Trainingsphase mit homomorph verschlüsselten Daten ein Modell trainiert werden kann, als auch in der Inferenzphase verschlüsselte Daten klassifiziert werden können [GLN12].

Je nach verwendeter HE-Ausprägung sind die möglichen Rechenoperationen eingeschränkt: *Partially HE* erlaubt nur eine einzige Operation auf verschlüsselten Daten, bspw. Multiplikation. Anwendungen mit *Somewhat HE* ermöglichen zwei verschiedene Rechenoperationen, bspw. Multiplikation und Addition. Die 2009 entwickelte *Fully HE* (FHE) [Gen09] schränkt die Rechenoperationen grundsätzlich nicht ein. Bei jeder Operation werden die verwendeten verschlüsselten Daten allerdings etwas verrauscht, weshalb nach einigen Operationen sog. *bootstrapping* (Neuberechnung und Wiederverschlüsselung) ausgeführt werden muss, um das Rauschen zurückzusetzen. Die Anwendungsmöglichkeiten von HE sind daher durch FHE zwar theoretisch gewachsen, aber das *bootstrapping* bringt einen Rechenaufwand mit sich, der in vielen Fällen nicht praktikabel ist. Eine Kompromisslösung stellt *Leveled HE* (LHE) dar, das

bootstrapping vermeidet und dennoch grundsätzlich alle Rechenoperationen erlaubt, wobei die Tiefe des zugrundeliegenden Schaltkreises für die Rechenoperationen im Vorhinein festegelegt werden muss.

Werden andere HE-Schemata als FHE oder LHE verwendet, müssen in Anwendungen für neuronale Netze die nicht-linearen Aktivierungsfunktionenen wie ReLU und Sigmoid (siehe Abschnitt 3.8) durch Polynome approximiert werden. Beispiele hierfür zeigen Hesamifard u. a., die einerseits ein neuronales Netz mit entsprechenden Ersatz-Aktivierungsfunktionen trainieren, und andererseits mit *CryptoDL* ein Framework für HE-gestützte Inferenz entwickelt haben. Eine Evaluation des Ansatzes zeigt lediglich minimale Performance-Einbußen durch das Ersetzen der Aktivierungsfunktionen und eine für viele Anwendungsfälle ausreichend schnelle HE-Inferenz von nahezu 164000 Vorhersagen pro Stunde [HTG16].

6.6 Trusted Execution Environments

Die in den vorhergehenden Abschnitten diskutierten Angriffs- und Verteidigungsmöglichkeiten setzen alle im Datenaustausch oder der Interaktion mit ML-Modellen an. Ein weiterer Angriffsvektor kann die Betriebssystemebene oder die Hardware betreffen, beispielsweise wenn im Rahmen von MLaaS (siehe Abschnitt 5) private Daten auf virtuellen Maschinen in einer Cloud verarbeitet werden. Ein bösartiger Cloud-Provider könnte den ausgeführten Code manipulieren oder sich mutwillig Zugriff auf die verarbeiteten Daten verschaffen. Durch *Trusted Execution Environments* (TEE, etwa "vertrauenswürdige Ausführungsumgebungen", auch: *Sichere Enklaven*) können die Authentizität von ausgeführtem Code, sowie die Integrität einer Laufzeitumgebung (inkl. CPU-Registern und des Arbeitsspeichers) und die Vertraulichkeit des ausgeführten Codes und der verwendeten Daten garantiert werden [SAB15]. Dafür sind spezielle Hardwarekomponenten nötig, die bspw. die Verschlüsselung von Bereichen im Prozessorenspeicher erlauben. Die Sicherheitsgarantien von TEEs basieren auf dieser spezieller Hardware. Es gibt allerdings auch Angriffe, die ebenjene Hardware komprimittieren [MIE17; Du+17].

Insbesondere in verteilten Szenarien können TEEs mit anderen ppML-Methoden kombiniert werden, um Vertrauen zu schaffen. So stellen bspw. Ohrimenko u. a. in [Ohr+16] einen Ansatz vor, in dem Teilnehmende gemeinsam auf einem Server mit einer TEE in Kombination mit Secure Multiparty Computation (siehe Abschnitt 6.4) ein ML-Modell trainieren können. Der auszuführende Code kann dabei gegenseitig überprüft werden und in geschützte Bereiche des Server-Arbeitsspeichers übertragen werden. Auch die Trainingsdaten können verschlüsselt in die sichere Enklave übertragen werden, wo im Anschluss an das Training das Modell in verschlüsselter Form vorliegt und heruntergeladen werden kann.

7 Machine Learning: Anwendungen im Gesundheitswesen

Als ein großer Bereich des alltäglichen Lebens mit maximaler Relevanz, in dem seit vielen Jahren systematisch und zunehmend digital Daten gesammelt werden, ist die Medizin ein vielversprechendes Anwendungsgebiet von ML-Technologien. Vor allem in folgenden Bereichen der Medizin kommt ML zum Einsatz [NK19; PSA18; Rav+16]:

Bildgebende Verfahren: Ultraschall-, CT-, MRT-, Röntgen- und andere bildgebende Verfahren sind ein wichtiger Bestandteil der Diagnostik in zahlreichen Disziplinen der Medizin. ML-Algorithmen können hierbei auf verschiedene Weisen unterstützen. Einerseits können entstehende Bilder für die behandelnden Ärzte vorverarbeitet werden, indem beispielsweise potenziell verdächtige bzw. relevante Teile der Bilder hervorgehoben oder ausgeschnitten werden. Dies sind klassische Anwendungsfälle für Muster- und Objekterkennungsalgorithmen. Andererseits können jene Algorithmen weiterhin dabei helfen, Bilder auszuwerten und Diagnosen zu erstellen. Insgesamt ist die Radiologie die medizinische Disziplin, in der ML-Algorithmen am meisten verbreitet sind [PSA18].

Durchgehende Gesundheitsanalysen: *Wearable Computers*, kurz *Wearables*, sind tragbare Computer, die während der Verwendung am Körper getragen werden. Beispiele sind Smartwatches und Smartbänder, die fortlaufend Daten der Vitalfunktionen (z.B. Pulsfrequenz, Bewegungsaktivitäten) der tragenden Person erheben, speichern und verarbeiten. Jene gesammelten Daten bergen in Kombination mit der Verarbeitung durch entsprechende (ML-)Algorithmen ein großes Potenzial für individuelle Gesundheits- und Ernährungsempfehlungen und die frühzeitige Erkennung von Auffälligkeiten. Andere *Wearables* wie beispielsweise Datenbrillen können mittels ML-getriebener 3D-Objekterkennung wertvolle Unterstützung im Alltag von Personen mit eingeschränkten Wahrnehmungsfähigkeiten leisten.

Ein Spezialfall der durchgehenden Vitalfunktionsanalyse sind Intensivstationen, in denen Patientinnen und Patienten in der Regel von mehreren Geräten überwacht werden. Hier können insbesondere rekurrente bzw. rückgekoppelte neuronale Netze (RNN) dabei helfen, realistische Modelle für das Auftreten von Notfällen wie Asthmaattacken zu entwickeln oder den Krankheitsfortschritt einzuschätzen [NK19].

Medizininformatik: Medizinische Gesundheitsakten einzelner Patientinnen und Patienten können große Datenmengen beinhalten, die ebenfalls viele Anwendungsmöglichkeiten bieten. Die Schwierigkeit uneinheitlicher (Freitext-)Einträge und Datenformate muss dabei überwunden werden. Insbesondere die frühzeitige Erkennung von Risikofaktoren bzw. ersten Anzeichen für bestimmte Krankheitsbilder (bspw. Diabetes, Krebs, Schizophrenie) kann durch umfassende ML-Auswertungen realisiert werden [Rav+16]. Ebenfalls existieren ML-Frameworks, die im Bereich der Onkologie anhand von Patientenakten individuelle Medikamentendosierungen errechnen [NK19].

Public Health: *Public Health* (dt. etwa „öffentliche Gesundheitsvorsorge“) ist ein im deutschsprachigen Raum noch wenig verankertes Konzept, das sich mit der bevölkerungsübergreifenden Krankheitsvorbeugung und Gesundheitsförderung beschäftigt. ML-Anwendungen können auch in diesem Bereich großen Nutzen generieren: Massenhafte ML-Datenauswertungen öffentlich zugänglicher Beiträge aus dem Nachrichtendienst Twitter wurden beispielsweise zur Erstellung von Gesundheitsstatistiken verwendet, ebenso wie Instagram-Beiträge für Analysen von Krankheiten wie Essstörungen und Mobilfunk-Metadaten für ML-Verhaltensanalysen [Rav+16].

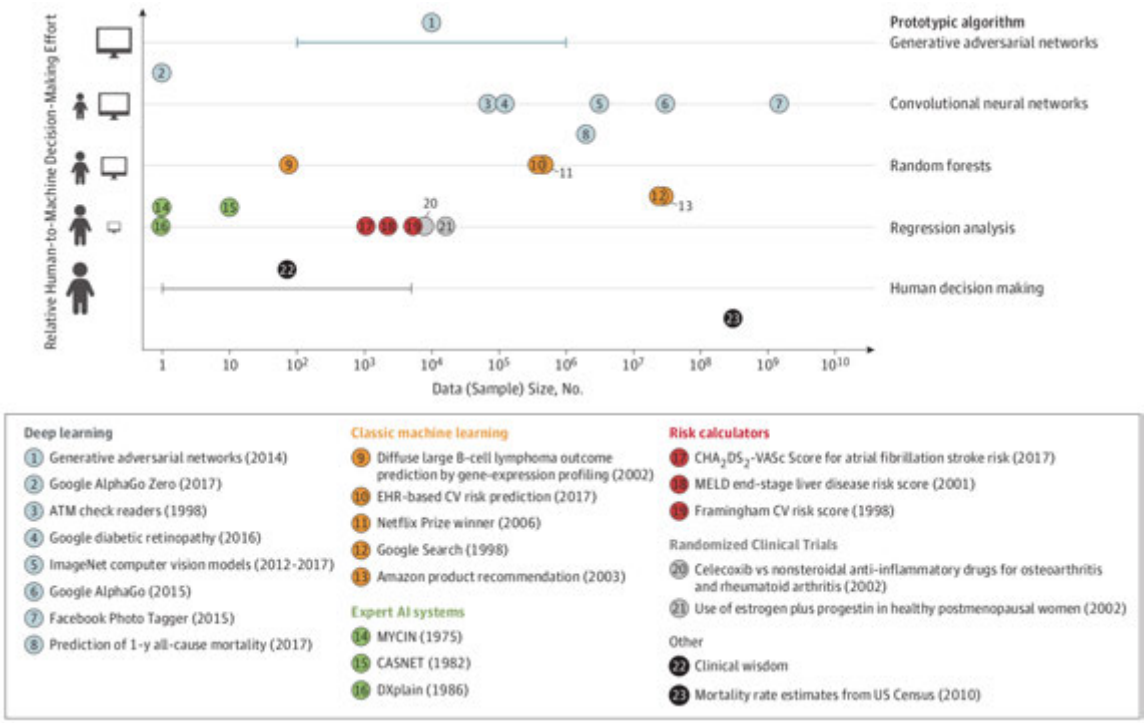

Abbildung 20: Kategorisierung von ML-Anwendungen und anderen Systemen aus [BK18].

Translationale Bioinformatik: Die translationale Forschung im Gesundheitswesen beschäftigt sich mit der Umsetzung von Forschungsergebnissen der (theoretischen) Gesundheitswissenschaften. Bioinformatische Anwendungen werden unter anderem für die Vorhersage biologischer Prozesse, für die Vermeidung von Krankheitsauftritten und für die individuelle Anpassung von Medikamenten für Patientinnen und Patienten eingesetzt. ML-Anwendungen können hier insbesondere bei der Verarbeitung von großen Datenmengen unterstützen [Rav+16].

7.1 Kategorisierung von medizinischen ML-Anwendungen

Aufgrund des großen Spektrums verschiedener ML-Anwendungen im Bereich der Medizin ist es schwierig, allgemeingültige Thesen aufzustellen. Die in diesem Abschnitt vorgestellten Systematiken sollen dabei helfen, Fallgruppen für die später aufgestellten Thesen zu bilden.

7.1.1 Kategorisierungen in der Literatur

Ein zentrales Kriterium für die Kategorisierung von ML-Algorithmen ist das Maß der menschlichen Einflussnahme. Beam und Kohane stellen in [BK18] eine Kategorisierung anhand zweier Achsen vor, die neben dem Maß menschlicher Beteiligung die Menge an Daten, die ein ML-Algorithmus in der Trainingsphase benötigt, berücksichtigt. Beispielhaft sortieren die Autoren 23 ML-Anwendungen in ihr Schema ein – von populären Algorithmen wie AlphaGo von Google über medizinische Risikoberechnungen bis hin zu vollkommen menschlicher „Weisheit“ (engl. *wisdom*; siehe Abbildung 20).

7.1.2 Trainingsdaten und Auswirkungen in der Inferenzphase

Als weitere Kriterien lässt sich einerseits die *Kritikalität von Trainingsdaten* nennen. Diese bezieht sich darauf, inwiefern Trainingsdaten privatsphärerelevante bzw. sensible Informationen

enthalten, die bspw. durch Angriffe auf die Privatsphäre von trainierten ML-Modellen (siehe Abschnitt 5) wiederhergestellt werden könnten.

Andererseits lohnt es sich, die *direkten Auswirkungen auf Betroffene während der Inferenzphase* zu betrachten. Hierfür ist der Einsatzkontext entscheidend und die Gewichtung von Entscheidungen des jeweiligen KI-Systems: Werden Entscheidungen ohne weitere manuelle Überprüfungen umgesetzt, kann dies ggf. direkte Konsequenzen für Betroffene mit sich bringen, von abgelehnten Krediten bis hin zur Gefährdung von Leib und Leben. Sind KI-Entscheidungen hingegen in einen Prozess eingebettet, sodass sie lediglich Experten in ihren (vom System unabhängigen) Handlungen unterstützen, kann der Einsatz eines KI-Systems ggf. als weniger kritisch gesehen werden.

In Abschnitt 7.3 werden beide Kriterien jeweils explizit diskutiert.

7.2 Regulierung

7.2.1 Behördliche Regulierung in den USA

In den USA reguliert die Behörde *Food and Drug Administration* (FDA) unter anderem den Markt für medizinische Geräte. Dies schließt explizit auch hardwareunabhängige Algorithmen (und somit potenziell auch ML-Anwendungen) ein, genannt „Software as a Medical Device“. Bestimmte Programme fallen seit 2016 allerdings *nicht* mehr unter die Aufsicht der FDA [Duk19, S. 13]: Einerseits muss es sich dabei um Software handeln, die keine medizinischen Bilder oder Signale von einem In-vitro-Diagnostikum oder Signale bzw. Muster von einem Signalerfassungssystem sammelt, verarbeitet oder analysiert. Andererseits darf die Software entweder nur eine einfache Aufgabe der Datenanalyse oder -darstellung übernehmen oder sie darf lediglich Empfehlungen an eine medizinische Fachkraft geben, die die Entscheidungsbasis für die Empfehlung transparent nachvollziehen kann und somit nicht nur aufgrund der Algorithmusausgabe handelt [Kon16, S. 1131].

7.2.2 Behördliche Regulierung in der EU

Vergleichbar mit der Regulierung in den USA gibt es in der EU die europäische Medizinprodukte-Verordnung (bzw. *Medical Device Regulation*, MDR), offiziell *Verordnung (EU) 2017/745*. Sie teilt Medizingeräte in vier Klassen I bis IV (geregelt in Anhang VIII) mit aufsteigendem Risiko für die Patienten und stellt an jede Klasse andere Anforderungen. Medizinprodukt-Software wird grundsätzlich der Klasse IIa zugeordnet, wenn sie Informationen liefert, „die zu Entscheidungen für diagnostische oder therapeutische Zwecke herangezogen werden“ (Verordnung (EU) 2017/745, Anhang VIII, Regel 11). Ausnahmen gibt es, falls diese Entscheidungen „den Tod oder eine irreversible Verschlechterung des Gesundheitszustands einer Person“ (Klasse III) bzw. „eine schwerwiegende Verschlechterung des Gesundheitszustands“ (Klasse IIb) bewirken können. Werden physiologische Prozesse per Software kontrolliert, wird diese der Klasse IIa zugeordnet, außer die Änderung der Parameter kann zu einer unmittelbaren Gefahr führen. In letztem Fall wird die Software in Klasse IIb eingeordnet. Andere Medizinprodukt-Software wird der Klasse I zugeordnet.

Für die Frage, ob Software-Anwendungen (und somit auch KI-Systeme) als Medizinprodukt gelten und unter dementsprechende Regulierung fallen, stellt die Europäische Kommission ein Dokument zur Orientierung zur Verfügung, die „Guidance on Qualification and Classification of Software in Regulation“ [Eur19]. So fallen beispielsweise Softwareanwendungen, die lediglich einfache Suchanfragen in der medizinischen Verwaltung ermöglichen, nicht unter die Medizinprodukte-Verordnung. Software, die medizinische Informationen verarbeitet, analysiert,

Fall-gruppe	Aufgabe des KI-Systems	Verwendete Daten	Kritikalität Trainingsdaten	Kritikalität Inferenz
F1	Molekülanalyse Arzneimittelforschung	Moleküldaten	unbedenklich	unbedenklich
F2	Alterseinschätzung	Portraitfotos	unbedenklich	eher unbedenklich
F3	Krankheits-Risikoscore App	Patientenakten	kritisch	eher unbedenklich
F4	Krankheits-Risikoscore mit Therapieempfehlung	Patientenakten	kritisch	kritisch
F5	Dosierung von Medikamenten	Patientenakten	unbedenklich	kritisch

Tabelle 2: Übersicht über die fiktiven Fallgruppen mit einer jeweils groben Einschätzung der datenschutzbezogenen Kritikalität.

erstellt oder verändert, gilt grundsätzlich dann als Medizinprodukt-Software, wenn die Erstellung oder die Veränderung der Informationen durch eine medizinische Zweckbestimmung bestimmt ist [Eur19, S. 6]. Softwareprodukte, die als Zubehör eines medizinischen Gerätes gelten oder ein medizinisches Gerät steuern oder beeinflussen, werden ebenfalls als Medizinprodukte eingeordnet. Ferner muss die Software individuellen Patienten helfen – bspw. im Gegensatz zu generischen Leitfäden, die nicht an individuelle Fälle gebunden sind [Eur19, S. 8-9].

Zur Regulierung von KI-Systemen im Speziellen hat die EU-Kommission vor Kurzem einen Gesetzesvorschlag vorgelegt, der zur Etablierung und Einhaltung einheitlicher Standards bezüglich des Datenschutz, digitaler Rechte und ethischer Grundlagen [Kom21]. Ein Beschluss ist nicht vor 2022 zu erwarten, da der Vorschlag zuvor vom Rat der EU-Staaten und vom EU-Parlament diskutiert, ggf. verändert und akzeptiert werden muss [Fan21].

7.3 Fallgruppen

In diesem Abschnitt werden die fünf fiktiven Fallgruppen F1–F5 exemplarisch skizziert, die als Grundlage für die Diskussion der rechtlichen Fragen in Abschnit 8 dienen sollen. In jeder Fallgruppe wird jeweils ein Szenario beschrieben, in dem ein KI-System eingesetzt wird. Die Eigenschaften der für die *Trainingsphase* relevanten Daten und die möglichen Auswirkungen der Vorhersagen in der *Inferenzphase* der zu den Fallgruppen gehörenden KI-Systeme sind dabei jeweils hervorgehoben. Eine Übersicht der Fallgruppen ist in Tabelle 2 gegeben.

(F1) Diese Fallgruppe ist in der Erforschung neuer Arzneimittel angesiedelt. Nach der Zielsetzung, ein neues Medikament gegen bestimmte Krankheiten oder Symptome zu entwickeln, müssen zunächst Millionen von Molekülen untersucht werden, um vielversprechende Kombinationen zu entdecken, die im Anschluss weiter optimiert werden. KI-Systeme sind für diese Untersuchung gut geeignet.

Trainingsdaten: Hierfür können öffentliche Datensätze mit entsprechenden Lizenzen verwendet werden [Ram+15].

Inferenz: Die Untersuchung von Molekülen betrifft nicht die Privatsphäre einzelner Personen. Die weitere Entwicklung eines neuen Arzneimittels ist streng reguliert und beinhaltet zahlreiche manuelle Überprüfungsschritte, sodass es sehr unwahrscheinlich ist, dass sich Fehler des KI-Systems in der Vorauswahl von geeigneten Molekülen bis hin zur Anwendung eines Medikamentes auswirken.

(F2) Dieses KI-System wurde von einer Digitalagentur mithilfe von öffentlich zugänglichen und entsprechend lizensierten Daten trainiert, die für tausende Personen jeweils ein Foto und das Alter der Person enthält. Nutzerinnen und Nutzer können in einer App eigene Bilder auf den Server der Agentur hochladen, um im Anschluss das KI-Modell zu konsultieren und als Ausgabe das (geschätzte) Alter einer abgebildeten Person erfahren.

Trainingsdaten: Die Daten sind veröffentlicht und dürfen unter entsprechender Lizenz verwendet werden.

Inferenz: Für die Verarbeitung des Fotos auf dem Server der Digitalagentur muss eine entsprechende Einwilligung der betroffenen Person gegeben werden. Die Entscheidung (geschätztes Alter) hat keine ernstzunehmenden Auswirkungen und ist daher unkritisch.

(F3) Das Modell in diesem KI-System wurde mit historischen Patientendaten von Personen trainiert, die an einer spezifischen Volkskrankheit leiden bzw. gelitten haben und vor der Verarbeitung ihrer Daten ihre Einwilligung gegeben haben. Das KI-System kann für eine Person anhand von Daten ihrer Patientenakte einen Risikoscore (von 0 bis 100) ermitteln, der das Risiko quantifiziert, mit der vorliegenden Symptomatik in den kommenden fünf Jahren mit einer deutlichen Verschlimmerung des Krankheitsverlaufs rechnen zu müssen. Patientinnen und Patienten nehmen freiwillig an dem Programm teil und erfahren ihren individuellen Risikoscore durch das Aufrufen einer App auf ihren Smartphones. Die Freischaltung für die Nutzung der App ist nur im Rahmen eines ärztlichen Gesprächs möglich, in dem die Eignung zur Teilnahme festgestellt werden muss. Erst dann kann die Übertragung von Patientendaten durch das medizinische Personal freigeschaltet werden. Eine textuelle Erklärung des Scores in der App hilft den Patientinnen und Patienten bei der Einordnung und gibt Handlungsempfehlungen (bspw. eine Empfehlung zu mehr sportlicher Betätigung). Bei einem auffallend schlechten Ergebnis wird Nutzerinnen und Nutzern das erneute Konsultieren ärztlichen Personals empfohlen.

Trainingsdaten: Für die Verarbeitung der Patientendaten liegen Einwilligungen für das Training des KI-Systems vor.

Inferenz: Teilnehmende Personen müssen auch hier ihre Einwilligung zur Verarbeitung ihrer Daten leisten. Die Konsequenzen einer Entscheidung des KI-Systems (der errechnete Risikoscore) sind begrenzt, da in der App lediglich leicht personalisierte Handlungsempfehlungen angezeigt werden. Die medizinischen Entscheidungen bleiben vollständig geschultem Personal überlassen.

(F4) Auch dieses KI-System basiert auf historischen Patientendaten und ist auf eine Volkskrankheit spezialisiert. Im Gegensatz zu dem vorhergehenden System besteht die Ausgabe allerdings aus Empfehlungen zur Therapie und Medikation von Patientinnen und Patienten, die anhand deren Symptomatik errechnet werden. Das System wird von ärztlicher Seite aus bedient, wo auch die Empfehlungen angezeigt werden. Das medizinische Personal hat

die Möglichkeit, die Empfehlungen des KI-Systems direkt zu übernehmen, kann sie ggf. aber auch ignorieren und andere Maßnahmen ergreifen.

Trainingsdaten: Siehe vorheriges KI-System in Fallgruppe F3.

Inferenz: Hierbei ist zu beachten, dass die Ausgabe des KI-Systems lediglich eine Empfehlung an das ärztliche Personal abgibt und diese nicht automatisch umgesetzt wird. Andernfalls würde dies einen klaren Verstoß gegen Art. 22 DSGVO darstellen. Dieser Thematik ist in Abschnitt 8.3 eine eigene These gewidmet.

(F5) Dieses fiktive KI-System kommt in der Patientenversorgung zum Einsatz. Ein Medizingerät wurde auf einem großen Datensatz anonymisierter Behandlungsdaten darauf trainiert, automatisch Dosierung für täglich verabreichte Medikamente anhand von aktuellen Vitaldaten der Betroffenen zu ermitteln. Die Medikamente werden durch das Gerät direkt vorbereitet und ausgegeben. Ein Eingriff von medizinischem Fachpersonal ist nicht vorgesehen.

Trainingsdaten: Dieses KI-System nutzt lediglich anonymisierte Trainingsdaten ohne Personenbezug.

Inferenz: Eine fehlerhafte Ausgabe von Medikamenten durch das KI-System kann negative Langzeitfolgen oder im schlimmsten Fall auch akute lebensbedrohliche Folgen besitzen.

8 Thesen

Auf den folgenden Seiten werden Thesen aufgestellt, die in acht Themengebiete gegliedert sind. Jedes Unterkapitel widmet sich je einem Themengebiet und beginnt mit einleitenden Hintergrundinformationen. Im Anschluss werden offene Fragen gestellt, die durch die Thesen beantwortet werden.

8.1 Rechtsgrundlage für die Verarbeitung

Hintergrund

Diese These beschäftigt sich mit der Rechtsgrundlage für die Verarbeitung von Daten, die in der Traingsphase eines maschinellen Lernverfahrens verwendet werden. Wenn diese Daten personenbezogen sind, muss eine Rechtsgrundlage für ihre Verarbeitung bestehen. Eine Möglichkeit hierfür ist die Einholung einer Einwilligung der betroffenen Person gemäß Art. 6 Abs. 1a) DSGVO bzw. Art. 7 DSGVO. Weiterhin sind aber auch andere Rechtsgrundlagen möglich, wie die Verarbeitung zu wissenschaftlichen Forschungszwecken.

Fragen

- Welche Rechtsgrundlagen können für den Einsatz eines ML-Verfahrens, speziell in der medizinischen Forschung, bestehen?
- Wie muss eine Einwilligung eines Betroffenen – u. a. eine konkrete und zulässige Zweckbeschreibung – formuliert sein, um eine rechtswirksame Datenverarbeitung zu ermöglichen?

These

Eine mögliche Rechtsgrundlage, die im Kontext medizinischer Forschung relevant sein könnte, ist in § 27 des Bundesdatenschutzgesetzes (BDSG) beschrieben. Hiernach können sensible Daten nach Art. 9 Abs. 1 der DSGVO auch ohne Einwilligung für Forschungszwecke verwendet werden, wenn die Verarbeitung erforderlich ist und die Interessen des Verantwortlichen die Interessen der betroffenen Person an einem Verarbeitungsausschluss erheblich überwiegen. In diesem Fall sind umfangreiche Maßnahmen zum Schutz der Betroffenen nach § 22 Absatz 2 des BDSG und Art. 89 DSGVO vorgesehen.

Eine weitere Möglichkeit besteht in der ausschließlichen Verwendung anonymisierter Daten, die nicht mehr unter den Geltungsbereich der DSGVO fallen. Gemäß Erwägungsgrund 26 sind hier verschiedene Faktoren wie etwa der Zeitaufwand einer möglichen Re-Identifizierung heranzuziehen, um zu bewerten, ob Daten als anonym angesehen werden können.

Wenn die Verarbeitung basierend auf der Einwilligung von Betroffenen geschieht, muss der Zweck der Verarbeitung eindeutig und klar verständlich definiert sein. Die Zwecksetzung muss dabei legitim und von inhaltlich ähnlichen Verarbeitungstätigkeiten klar getrennt sein (siehe Art. 7 Abs. 2 DSGVO). Ferner muss im Zweifel nachweisbar sein, dass die Zweckbindung der Datenverarbeitung über alle organisatorischen und infrastrukturellen Grenzen der verantwortlichen Organisation hinweg implementiert ist. Wenn es sich bei der Verarbeitung um das Trainieren von KI-Modellen handelt, ist dieser Vorgang für Laien verständlich zu erklären. Außerdem sind Schutzmaßnahmen zu benennen, die getroffen wurden, um Datenmissbrauch (bspw. Angriffe auf KI-Modelle) zu verhindern. Auch die Schutzmaßnahmen sind verständlich und in einer klaren und einfachen Sprache zu erklären.

8.2 Datenschutz-Folgenabschätzung

Hintergrund

Art. 35 DGSVO fordert „eine Abschätzung der Folgen der vorgesehenen Verarbeitungsvorgänge für den Schutz personenbezogener Daten“, insbesondere wenn neue Technologien verwendet werden und wenn „aufgrund der Art, des Umfangs, der Umstände und der Zwecke der Verarbeitung voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen zur Folge“ besteht. In Art. 35 Abs. 3 lit a–c werden Szenarien genannt, für die eine Datenschutz-Folgenabschätzung (DSFA) insbesondere erforderlich ist.

Für die Anfertigung einer DSFA kann sich die verantwortliche Stelle an den Schutzzielen des Standard-Datenschutzmodells orientieren [Der19]. Grundsätzlich muss dabei das System zunächst detailliert beschrieben werden, um es im Anschluss auf Risiken zu untersuchen. Die dabei identifizierten Risiken müssen im Anschluss bewertet werden, woraufhin Handlungsmöglichkeiten abgeleitet werden, die möglichst auch angewandt werden sollten. Die Anfertigung einer DSFA ist somit in der Regel ein iterativer Prozess, der auch nach Inbetriebnahme des Systems fortgeführt werden muss.

Fragen

- Welche Kriterien müssen erfüllt sein, damit eine DSFA für ein KI-System angefertigt werden muss?

These

Angesichts des großen aktuellen gesellschaftlichen Diskurses und der teils unklaren Rechtsauslegung sind KI-Systeme überwiegend als „neue Technologie“ zu bewerten. Damit wird bereits ein Kriterium für die Notwendigkeit der Anfertigung einer DSFA erfüllt. Mit dem oft datenintensiven Training von KI-Algorithmen wird weiterhin der Bestand der „systematischen“ und „umfangreichen“ Verarbeitung von vielen KI-Systemen erfüllt. Werden dabei beispielsweise besonders sensible Daten gemäß Art. 9 Abs. 1 DSGVO verarbeitet, dürfte die Erstellung einer DSFA in vielen Fällen unumgänglich sein (siehe auch [Art17, S. 9–11]).

Eine Schwierigkeit, die sich durch die teils schwerwiegenden Konsequenzen durch den Einsatz von ML-Technologien ergibt, ist, dass die DSGVO großteils *individuelle* Betroffenenrechte und Pflichten auf Seiten der verarbeitenden Stellen betrachtet. Allerdings sind – je nach Anwendungsbiet – mögliche Auswirkungen auf *gesamtgesellschaftliche* Rechte und Freiheiten in einer Folgenabschätzung oft mindestens genauso relevant und untersuchenswert. So plädiert beispielsweise auch die Datenethikkommission in ihrem Gutachten je nach Schädigungspotenzial eines KI-Systems für eine Abschätzung der „Risiken für Selbstbestimmung, Privatheit, körperliche Unversehrtheit, persönliche Integrität sowie für Vermögen, Eigentum, und Gleichbehandlung“ [Dat19, S. 188]. Ferner sind zahlreiche zivilgesellschaftliche und öffentliche Organisationen zurzeit damit beschäftigt, geeignete Rahmenwerke und umfassende Verfahren zu entwickeln.

Antithese

Eine pauschale Einstufung von KI-Systemen als „neue Techologien“ ist nicht haltbar. Angesichts der großen Vielfalt an Algorithmen unter dem Deckmantel des Maschinellen Lernens existieren auch zahlreiche etablierte und erprobte Verfahren, die wenig mit der teilweise intransparenten Natur von bspw. Neuronalen Netzen gemein haben. Wird eine Technologie nicht als neu

eingestuft, trifft eines der Kriterien für die Anfertigung einer DSFA nach Art. 35 Abs. 1 nicht mehr zu.

Ferner sind auch die weiteren Kriterien, die eine DSFA notwendig machen, differenziert zu betrachten. Insbesondere bei Systemen, mithilfe derer entweder keine personenbezogene Daten verarbeitet werden (wie in Fallgruppe F1), oder deren Verarbeitungsvorgänge voraussichtlich keine gravierenden Auswirkungen für Betroffene nach sich ziehen können, ist das „hohe Risiko für die Rechte und Freiheiten natürlicher Personen“ in Art. 35 Abs. 1 mutmaßlich nicht gegeben. Somit muss auch für KI-Systeme jeweils individuell geprüft werden, ob die Anfertigung einer DSFA notwendig ist.

Fallgruppen

(F1) Da in dieser Fallgruppe keine personenbezogenen Daten verarbeitet werden, muss auch keine DSFA angefertigt werden.

(F2) Der Einsatz des KI-Systems in Fallgruppe F2 birgt voraussichtlich keine Risiken für die Rechte und Freiheiten natürlicher Personen, daher ist gem. Art. 35 Abs. 1 DSGVO auch keine DSFA anzufertigen.

(F3) Auch die Verarbeitung in Fallgruppe F3 zieht keine direkten Konsequenzen für Betroffene nach sich, sondern resultiert lediglich in eher allgemeingültige Handlungsempfehlungen. Daher ist analog zu Fallgruppe F2 keine DSFA anzufertigen.

(F4) Da in Fallgruppe F4 medizinische Maßnahmen vorgeschlagen werden, muss für die Einschränkung von Freiheiten Betroffener und somit die Notwendigkeit einer DSFA argumentiert werden.

(F5) Die potenziellen Konsequenzen für Betroffene im Rahmen einer automatisierten Verarbeitung sensibler Daten sind erheblich, sodass die Anfertigung einer DSFA notwendig ist.

8.3 Automatisierte Entscheidungen

Hintergrund

Laut Art. 22 Abs. 1 DSGVO dürfen Betroffene „nicht einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung“ mit rechtlicher oder vergleichbar beeinträchtigender Wirkung unterworfen werden. Es gibt allerdings Ausnahmen (Art. 22 Abs. 2 DSGVO), die die ausdrückliche Einwilligung betroffener Personen mit einschließen. Dies gilt auch für die Verarbeitung „besondere[r] Kategorien personenbezogener Daten“ (Art. 22 Abs. 4 DSGVO), in die Betroffene ausdrücklich einwilligen können (Art. 9 Abs. 2 lit. a DSGVO).

Art. 22 Abs. 3 DSGVO stellt ferner fest, dass Betroffene „mindestens das Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen“ haben. Dieses Recht wird hier nicht im Kontext der ML-Trainingsphase betrachtet, sondern für die anschließende Inferenz-Phase, in der Anfragen an das ML-System gestellt werden können. Diese Anfragen können auch personenbezogene Daten enthalten, auf deren Basis der KI-Algorithmus durch Konsultation des trainierten Modells Aussagen trifft.

Fragen

- Inwieweit dürfen ML-Systeme rechtlich autonom handeln und Entscheidungen treffen? Gibt es hier maßgebliche Unterschiede zwischen der Verarbeitung von gewöhnlichen, personenbezogenen Daten und besonderen Daten gem. Art. 9 Abs. 1 DSGVO?
- Unter welchen Umständen muss das Recht auf Erwirkung des Eingreifens und der Anfechtung der Entscheidung gewährt werden? Wie können Personen bei der verantwortlichen Stelle (rechtswirksam gem. Art. 22 Abs. 3 DSGVO) in ein ML-System „eingreifen“?

These

Die Rechtsauslegung dürfte vor allem von zwei Fragen abhängig sein:

1. Sind durch die Entscheidung des KI-Systems weitreichende, praktische Konsequenzen für Betroffene zu befürchten?
2. Falls die erste Frage bejaht werden kann: Werden die Empfehlungen eines KI-Systems direkt umgesetzt, oder wird sichergestellt, dass sie zuvor von professionell geschultem Personal inhaltlich geprüft wird?

Automatisierte Entscheidungen dürften rechtlich unproblematisch sein, solange die erste Frage verneint werden kann. Die zweite Frage bezieht sich auf die Definition in Art. 22 Abs. 1 DSGVO der „ausschließlich automatische[n]“ Entscheidungen: Nur wenn fachliche Experten eine algorithmische Empfehlung umsetzen, deren Grundlagen sie nachvollziehen können und inhaltlich prüfen, entfällt der Status der ausschließlichen Automatisierung. Dies gilt sowohl für gewöhnliche personenbezogene Daten, als auch für Daten besonderer Kategorien gem. Art. 9 Abs. 1 DSGVO.

Eine strenge Interpretation der Verordnung würde hingegen einen Vorgang, in dem eine Fachkraft eine algorithmische Handlungsempfehlung durch einfache Bestätigung umsetzt, als nicht vollständig automatisiert einstufen – auch wenn sich der menschliche Anteil auf einen einzigen Klick beschränkt. Es lässt sich somit argumentieren, dass eine derart strenge Auslegung die Norm letztlich unbrauchbar mache [HN18, S. 53]. Erst wenn sich ein Mensch *inhaltlich* mit der Entscheidung auseinandergesetzt hat, entfällt der Status der ausschließlichen Automatisierung.

Antithese

Eine strenge Auslegung der DSGVO verbietet den Einsatz vollständig autonomer Systeme, die ohne menschliche Eingriffe funktionieren, wenn sie Personen erheblich beeinträchtigt. Dies würde folglich bestimmte Technologien wie das in Fallgruppe F5 beschriebene Gerät zur Medikamentenverabreichung verhindern. In der Fallgruppe ließe sich die Möglichkeit menschlichem Eingreifens integrieren. Aber in möglichen Zukunftstechnologien wie etwa bei autonomen ML-basierten Operationsrobotern, wenn etwa die Verarbeitung schneller geschehen muss als ein Mensch reagieren könnte, wäre die Eingriffsmöglichkeit praktisch nicht mehr gegeben. Hierdurch könnten Erkrankte Vorteile dieser Technologien, wie etwa eine potentiell deutlich höhere Präzision gegenüber menschlichen Behandlern, nicht nutzen. Auch wenn entsprechende Systeme bestmöglich reguliert und überprüft werden müssen, so sollte eine rechtlicher Rahmen für ihren Einsatz diskutiert werden.

Fallgruppen

Ausschlaggebend für die Bewertung der Fallgruppen dürften vor allem die möglichen Konsequenzen einer von KI-Systemen gefällten Entscheidung sein:

(F1) Da in dieser Fallgruppe keine Personen involviert sind, ist die in der These fokussierte Problematik unerheblich.

(F2) Während es sich bei der Datenverarbeitung in Fallgruppe F2 zwar um personenbezogene Daten handeln kann, sind die Auswirkungen der KI-Entscheidung (Kategorisierung des Alters) im privaten Anwendungsumfeld minimal.

(F3) In Fallgruppe F3 werden zwar Gesundheitsdaten verarbeitet, die gemäß Art. 9 Abs. 1 DSGVO unter speziellem Schutz stehen, allerdings sind die Konsequenzen auch in diesem Fall noch überschaubar. Im schlimmsten Fall wird das Risiko für einen schweren Krankheitsverlauf für eine betroffene Person als zu niedrig eingeschätzt. Da die Verwendung der Applikation, wie in Fallgruppe F3 beschrieben, allerdings auf Freiwilligkeit basiert, dürfte diese Fehlentscheidung aber ohnehin während der nächsten ärztlichen Untersuchung auffallen. Insofern ist der „Handlungsspielraum“ des KI-Systems insoweit eingeschränkt, dass keine deutlichen praktischen Nachteile bei betroffenen Personen entstehen können.

(F4) In Fallgruppe F4 muss hingegen sichergestellt werden, dass sich das ärztliche Personal auch unabhängig von der klaren Empfehlungen des KI-Systems weiterhin mit der Patientenakte beschäftigt. Ansonsten besteht die Gefahr, dass Empfehlungen aus dem System übernommen werden, ohne sich *inhaltlich* damit auseinander zu setzen.

(F5) Fallgruppe F5 birgt das Risiko ernsthafter Langzeitfolgen bis zum Tod einer Person und ist daher mit großer Wahrscheinlichkeit unzulässig. In diesem Szenario ist die Verarbeitung personenbezogener Daten eher beschränkt, sodass wahrscheinlich eher die Rechtmäßigkeit im Rahmen der Medizinprodukte-Verordnung betrachtet werden muss.

8.4 Personenbezogene Daten: Angriffe

Hintergrund

Es existieren eine Reihe von Angriffen auf KI-Modelle, mit deren Hilfe personenbezogene Daten aus dem Trainingsdatensatz extrahiert werden können. Für diese These spielen insbesondere die beiden Angriffe *Model Inversion* und *Membership Inference* eine Rolle. Während *Model Inversion* darauf abzielt, durch sukzessives Raten die ursprünglichen Trainingsdaten zu rekonstruieren, kann durch *Membership Inference* verifiziert werden, dass bestimmte Datenpunkte im Trainingsdatensatz vorhanden waren (siehe Abschnitt 5).

Fragen

- Inwiefern sind KI-Modelle (Speichern von Modellparametern), die mit (personenbezogenen) Daten angelernt wurden, selbst als personenbezogene Daten zu bewerten? Diese Frage stellt sich insbesondere vor dem Hintergrund von Privatsphäre-Angriffe wie *Model Inversion* und *Membership Inference*.

These

Während der Trainingsphase werden für die Aufgabe eines KI-Systems relevante Muster aus den Trainingsdaten extrahiert. Handelt es sich dabei um personenbezogene Daten, können neben abstrakten Mustern aber auch sensible Informationen im ML-Modell gespeichert werden. Der Erfolg von Angriffen wie *Model Inversion* oder *Membership Inference* zeigt dies: Wenn durch Model Inversion Trainingsdaten wiederhergestellt werden können, sind offenbar Repräsentationen der Daten im Modell gespeichert. Je nach Genauigkeit des Angriffsergebnisses kann dies große Einschränkungen der Privatsphäre Betroffener bedeuten. Handelt es sich beispielsweise um ein KI-Modell zur Erkennung von Straftätern und ein identifizierbares Foto eines Täters kann aus dem Modell wiederhergestellt werden, kann dies direkte Auswirkungen auf Betroffene haben. Ähnlich verhält es sich bei erfolgreichen *Membership-Inference*-Angriffen: Hier kann auf direktem Wege verifiziert werden, dass bspw. eine Person Teil eines Trainingsdatensatzes für eine Straftäter-Erkennungssoftware war.

Somit kann argumentiert werden, dass auch ML-Modelle selbst personenbezogene Daten sind, die durch Privatsphäreangriffe extrahiert und verwendet werden können. Dies würde unter anderem bedeuten, dass jede Anfrage an das Modell als Verarbeitung personenbezogener Daten zu verstehen ist und die Betroffenenrechte der DSGVO (insbesondere Art. 15-18) auch im Zusammenhang mit KI-Modellen gewahrt werden müssen.

Verantwortliche, die einen KI-Algorithmus Externen zur Verfügung stellen, müssen durch möglichst umfangreiche Analysen sicherstellen, dass sie gegen bekannte (bspw. die oben genannten) Angriffe Sicherheitsvorkehrungen getroffen haben. Auch die Anfertigung einer Datenschutz-Folgenabschätzung sollte bei kritischen Anwendungen immer in Betracht gezogen werden und häufig auch notwendig sein, siehe These 8.2. Diese Vorkehrungen müssen ausreichend dokumentiert werden.

Antithese

Wie in Abschnitt 5 geschildert, sind Privatsphäreangriffe nur unter bestimmten Umständen erfolgreich. Insbesondere *Model Inversion* scheint nur dann gut zu funktionieren, wenn die Daten einzelner Klassen in Trainingsdaten für eine Klassifizierungsaufgabe nicht allzu divers sind, da eine Art Durchschnitt der Daten einer Klasse rekonstruiert wird. Die Rekonstruktion sensibler Daten ist also keineswegs trivial und in vielen Fällen nicht möglich.

Auch ein *Membership-Inference*-Angriff ist in der Regel nur mit hohem Aufwand durchzuführen, da für die Entwicklung des Angriffalgorithmus eine große Menge an Daten notwendig ist, die den Trainingsdaten des anzugreifenden Zielmodells möglichst ähnlich sein sollten. Neben der Beschaffenheit der Trainingsdaten ist es außerdem hilfreich für Angreifende, die Architektur und Beschaffenheit des anzugreifenden Modells in Erfahrung zu bringen. Haben Angreifende nur Black-Box Zugriff und nur wenig Wissen über die Modellinterna und Trainingsdaten des Zielmodells, ist auch ein erfolgreicher *Membership Inference* Angriff nahezu unmöglich. Modelle können also unter Umständen gegen Privatsphäreangriffe geschützt werden, indem technische Spezifikationen geheimgehalten werden.

Alternativ dazu kann *Privacy-Preserving ML* (siehe Abschnitt 6) zur technischen Absicherung eines Modells eingesetzt werden, indem bspw. durch *Differential Privacy* Rückschlüsse auf Individuen in einem Trainingsdatensatz verhindert werden. Angriffe wie *Membership Inference* werden dadurch erschwert, wenn nicht sogar unmöglich gemacht.

Insgesamt kann auch mit der Verhältnismäßigkeit von Schutzbedarfen und ggf. notwendigen Schutzmaßnahmen argumentiert werden. Ist die Extraktion von sehr sensiblen Daten aus einem Modell möglich oder impliziert der Einsatzkontext eines KI-Systems große Risiken für

Betroffene (bspw. in der Straftätererkennung), ist die Notwendigkeit von Schutzmaßnahmen einfacher gerechtfertigt als in unkritischen Systemen, wie auch die untenstehenden Fallgruppen demonstrieren.

Fallgruppen

(F1) Da hier keine personenbezogenen Daten zum Einsatz kommen, kann auch das Modell nicht als personenbezogenes Datum gesehen werden.

(F2) Da die Trainingsdaten aus einer öffentlich zugänglichen Datenbank stammen, sind Privatsphäreangriffe auf dieses KI-System unkritisch.

(F3) ML-Modelle des Systems aus Fallgruppe F3 werden mit Gesundheitsdaten trainiert. Je nach Ausführung könnten diese rekonstruiert werden, was schwerwiegende Folgen für die Privatsphäre Betroffener bedeuten kann. Das Modell sollte gut gegen Angriffe geschützt werden.

(F4) Analog zu Fallgruppe F3 sollte auch das System in Fallgruppe F4 gut geschützt werden, da personenbezogene Daten aus dem Modell rekonstruiert werden könnten.

(F5) Die Trainingsdaten für das KI-System in Fallgruppe F5 lagen anonymisiert vor. Eine Rekonstruktion von personenbeziehbaren Daten aus einem Modell, das mit anonymisierten Daten trainiert wurde, ist durch beschriebene Angriffe schwer vorstellbar.

8.5 Rechtssichere technische und organisatorische Schutzmaßnahmen

Hintergrund

Wie in der These zur Absicherung gegen Privatsphäreangriffe angedeutet (siehe Abschnitt 8.4), können für den Einsatz von KI-Systemen Schutzmaßnahmen erforderlich sein. Die DSGVO schreibt in Art. 25 (Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen) den Einsatz geeigneter Maßnahmen zur wirksamen Umsetzung von Datenschutzgrundsätzen wie etwa der Datenminimierung vor. Art. 32 (Sicherheit der Verarbeitung) fordert vergleichbare Maßnahmen für ein angemessenes Schutzniveau der Verarbeitung. Jeweils sind hierbei unter anderem der Stand der Technik, Implementierungskosten und auch die Eintrittswahrscheinlichkeiten und Auswirkungen von bestehenden Risiken zu bewerten. Diese beiden Artikel beschreiben die in der Informatik bekannten Prinzipien von *Privacy by Design (PbD)* und *Security by Design (SbD)*.

Diese These beleuchtet die Bedeutung von PbD und SbD im Kontext maschineller Lernverfahren. Weiterhin wird die Rechtssicherheit des Einsatzes der in Abschnitt 6 vorgestellten Techniken für privatsphärefreundliches ML zur Umsetzung der oben genannten Prinzipien untersucht. Hierbei sind verschiedene Aspekte zu beachten, die die Notwendigkeit, die Art und den Umfang der Schutzmaßnahmen maßgeblich beeinflussen:

- Umfang der Verarbeitung personenbezogener Daten durch das KI-System.
- Architektur des ML-Verfahrens (zentral oder dezentral wie bspw. Federated Learning, Abschnitt 6.3)
- Einsatzgebiet und Kontext des KI-Systems (Welche Angriffe sind möglich oder gar zu erwarten?)

- Welche Phase des KI-Algorithmus muss geschützt werden – Training oder Inferenz?

In Abhängigkeit von diesen Faktoren können folgende Fragen gestellt werden:

Fragen

- Welche Maßnahmen (im Speziellen ppML-Maßnahmen) können als Stand der Technik angesehen und damit vorausgesetzt werden?
- Welches Angreifermodell muss grundsätzlich berücksichtigt werden, wenn Schutzmaßnahmen erforderlich sind (siehe bspw. Abschnitt 6.4)?

These

Die Verfahren aus Kapitel 6 sind in unterschiedlichem Maße als Stand der Technik anzusehen. Methoden der Pseudonymisierung und Anonymisierung werden seit vielen Jahren erforscht und auch praktisch eingesetzt. Ihr Einsatz kann daher vorausgesetzt werden, wenn es im Einsatzkontext sinnvoll ist. Der Einsatz von Differential Privacy und Federated Learning kann in manchen Szenarien sinnvoll sein und auch hier liegen bereits erste verfügbare Implementierungen für den produktiven Einsatz vor[3]. Nichtdestotrotz handelt es sich noch um relativ neue Verfahren, die in der Forschung aktiv untersucht werden. Eine Einordnung als dem Stand der Technik entsprechend ist hier wohl noch verfrüht. Secure Multiparty Computation und Homomorphe Verschlüsselung sind aktueller Forschungsgegenstand und kaum für den produktiven Einsatz geeignet. Gerade Fragen der Performanz lassen es fraglich erscheinen, ob dies jemals für alle Szenarien der Fall sein wird. Unabhängig davon sind die beiden Verfahren derzeit kaum als Stand der Technik anzusehen.

Neben diesen technischen Maßnahmen sind auch organisatorische Maßnahmen im Rahmen von ML-Verfahren zu bedenken. Hierunter können beispielsweise das Entfernen fachlich nicht notwendiger Features in den Trainingsdaten (Datenminimierung), das Löschen nicht mehr benötigter Trainingsdaten oder die Zugriffsbeschränkung auf sensible Trainingsdaten auch für Beschäftigte innerhalb der verantwortlichen Stelle fallen. Um Privatsphäreangriffe zu erschweren, kann weiterhin sichergestellt werden, dass ein ML-Modell nur als Black-Box erreichbar ist und dass – je nach Anwendungsfall – die Ausgabeinformationen minimiert wurden (bspw. in Klassifizierungsausgaben: Ausgabe der Klasse mit der höchsten errechneten Wahrscheinlichkeit anstatt der Ausgabe von Klassenzugehörigkeits-Wahrscheinlichkeiten).

Antithese

Viele der in Kapitel 6 vorgestellten Verfahren sind Gegenstand aktueller Forschung und derzeit höchstens für Fachleute einsetzbar. Dem verbreiteten Ansatz solcher Maßnahmen stehen derzeit noch praktische Gründe wie eine nicht ausreichende Performance entgegen. Häufig liegen weiterhin keine verfügbaren oder für den produktiven Einsatz geeigneten Implementierungen vor. In manchen Fällen führt der Einsatz der Techniken auch zu einem erhöhten Datenbedarf, der lediglich in wenigen Problemfeldern, aber nicht allgemeingültig erbracht werden kann. Die Angemessenheit und Verhältnismäßigkeit der Maßnahmen ist immer im Einzelfall zu betrachten, aber derzeit ist der verpflichtende Einsatz der hier erwähnten Maßnahmen für Verantwortliche im Allgemeinen noch als nicht verhältnismäßig anzusehen.

[3] *TensorFlow Federated* (https://github.com/tensorflow/federated) ist ein Framework, das Federated Learning auf Basis des bekannten ML-Frameworks TensorFlow anbietet. *Open Differential Privacy* (https://github.com/opendp) bietet verschiedene Schnittstellen für die Implementierung von DP an.

Fallgruppen

Da passende Schutzmaßnahmen sehr abhängig von Kontext und konkreter Umsetzung der jeweiligen KI-Systeme sind, wird für diese These nicht auf die abstrakten Beispielsysteme der Fallgruppen eingegangen.

8.6 Transparenz

Hintergrund

Gemäß Art. 12 DSGVO haben Verantwortliche eine möglichst weitgehende Offenbarungspflicht gegenüber den Betroffenen. Letztere haben das Recht auf einen transparenten Einblick in die Verarbeitung ihrer Daten, in „präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache“ (Art. 12 Abs. 1 DSGVO). Insbesondere bei einer automatisierten Entscheidungsfindung einschließlich Profiling (Art. 4, Abs. 4, Art. 22 Abs. 1 und 4 DSGVO; Erwägungsgrund 71) müssen unter anderem „aussagekräftige Informationen über die involvierte Logik“ zur Verfügung gestellt werden. Umstritten ist, wie die Verantwortlichen diesen Anforderungen bei der Anwendung von KI-Systemen gerecht werden können.

Es lässt sich zunächst feststellen, dass verschiedene technische Werkzeuge existieren, die das Erklären von Entscheidungsfindungen von KI-Algorithmen unterstützen können. Zahlreich sind vor allem *lokale* Erklärungsmodelle, die jeweils die wichtigsten Einflussfaktoren für *eine* Entscheidung offenlegen (für Beispiele und detaillierte Erklärungen siehe Abschnitt 4). Auch wenn dabei der Entscheidungsprozess an sich nicht vollständig erklärt wird, bieten diese Methoden wichtige Einblicke. Dadurch wird einerseits Transparenz gewährt (*„No more information is needed under the law.“* [Dos+17, S. 14]), andererseits bleiben die Interna des Algorithmus weitgehend versteckt, die oft als Geschäftsgeheimnis verstanden werden (beispielsweise im BGH-Urteil vom 28.01.2014, Az. VI ZR 156/13, damals auf Basis des Bundesdatenschutzgesetzes).

Neben technischen Werkzeugen gibt es weiterhin zahlreiche Empfehlungen und semantische Rahmenwerke für das Erklären von KI-Entscheidungen. Beispielsweise hat die britische Datenschutzbehörde *ICO* gemeinsam mit dem *Alan Turing Institute* Anfang des Jahres eine detaillierte informelle Richtlinie mit praktischen Handreichungen veröffentlicht [IT20]. Eine finale Version der Richtlinie, die die im Rahmen des Konsultationsverfahren erhaltenen Kommentare berücksichtigt, steht noch aus.

Neben Ansätzen der erklärbaren KI kann für die Erklärung einer KI-Entscheidung auch die Herausgabe von Trainingsdaten in Erwägung gezogen werden [HN18, S. 59]. Dies ermöglicht es einerseits, das Zustandekommen eines KI-Modells nachzuvollziehen, andererseits können dadurch mögliche Voreingenommenheiten (engl. *bias*) des Modells besser erklärt werden (siehe *Diskriminierung* in Erwägungsgrund 75 der DSGVO).

Fragen

- Wie weitreichend ist die Informationspflicht für Verantwortliche gemäß Art. 13/14 DSGVO? Inwiefern muss das Zustandekommen einer KI-Entscheidung erklärt werden?
- Kann sich der Auskunftsanspruch von Betroffenen auf die Trainingsdaten erstrecken? Wie wird das Recht auf Auskunft mit dem Recht auf Privatheit anderer Betroffener abgewogen?

- Wie beziehungsweise in welchem Rahmen kann die Verhinderung von Diskriminierung ohne die Veröffentlichung von Trainingsdaten bestmöglich und rechtssicher belegt werden?

These

Der Einsatz transparenzschaffender Hilfsmittel kann, sofern es im Kontext des jeweiligen KI-Systems mit vertretbarem Aufwand verbunden ist, von der verantwortlichen Stelle erwartet werden (siehe auch [FLS20]). Gemäß den Artikeln 13-15 der DSGVO bezieht sich die Informationspflicht bzw. das Auskunftsrecht im Falle einer automatisierten Entscheidungsfindung einschließlich Profiling auch auf aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen der Verarbeitung. Hier ist der Einsatz geeigneter transparenzschaffender Verfahren also verpflichtend.

Das Argument, einen KI-Algorithmus angesichts seiner Kompliziertheit nicht erklären zu können, ist nicht überzeugend: Mindestens Kontext, Herangehensweise und Datengrundlage sollten stets von den Verantwortlichen erklärt werden können (beispielsweise mithilfe bestehender Rahmenwerke [Geb+18; Mit+19]). Weiterhin kann es sinnvoll und zweckdienlich sein, die für das Training ausgewählten Features ebenfalls offenzulegen.

Die Herausgabe von Trainingsdaten als Erklärung für eine automatisierte Entscheidungsfindung ist unverhältnismäßig. Die Trainingsdaten lassen tiefe Einblicke in die Hintergründe einer ML-Entscheidung zu, jedoch kann es sich hierbei wiederum um personenbezogenen Daten anderer Betroffener handeln, die die verantwortliche Stelle schützen muss. Diese Auslegung ist beispielsweise auch in Bezug auf das Auskunftsrecht in der DSGVO verankert. Nach Art. 15 Abs. 4 darf das Recht auf Erhalt einer Kopie die Rechte und Freiheiten anderer Personen nicht beeinträchtigen.

Alternative Maßnahmen zur Transparenzgewinnung (siehe Abschnitt 4) sind somit zu bevorzugen. Um die Diskriminierungsfreiheit eines Algorithmus zu belegen, kann eine datenverarbeitende Stelle in Erwägung ziehen, Ausgabeverteilungen für verschiedene Eingaben des Algorithmus zu veröffentlichen [Hac18, S. 27].

Antithese

Die Pflicht zum Einsatz geeigneter transparenzschaffender Maßnahmen ist nicht immer vorhanden. Die in der DSGVO aus einer automatisierten Entscheidungsfindung einschließlich Profiling folgende Transparezpflicht bezieht sich lediglich auf Verfahren, die rechtliche Wirkung für die betroffene Person entfalten oder sie in ähnlicher Weise erheblich beeinträchtigen. Sind diese Auswirkungen nicht gegeben, so entfällt auch die Pflicht zum Einsatz geeigneter Maßnahmen.

Ein potentieller Zielkonflikt besteht zwischen dem Einsatz transparenzschaffener Maßnahmen und dem Schutz geistigen Eigentums kommerzieller Anbieter entsprechender Systeme. Hier muss im Einzelfall entschieden werden, welche Angaben in welchem Detailgrad im Sinne einer Transparenzpflicht öffentlich gemacht werden müssen.

Fallgruppen

Für die fokussierten Fallgruppen ergeben sich in Bezug auf Transparenzpflichten folgende Einschätzungen:

(F1) In diesem Szenario entstehen durch die DSGVO keine Transparenzpflichten. Für die Wirkstoffentwicklung mag das Verstehen des Verfahrens entscheidend sein. Da jedoch keine personenbeziehbaren Daten verarbeitet werden, besteht auch keine Informationspflicht.

(F2) Da aus dem System in dieser Fallgruppe keinerlei Folgen entstehen, kann auch der Einsatz transparenzschaffender Maßnahmen als nicht notwendig angesehen werden.

(F3) Ein Risikoscore kann als eine Art des Profilings angesehen werden. Selbst wenn der Score in diesem Fallbeispiel keine direkten Auswirkungen auf die betroffene Person hat, sollten trotzdem Informationen über die involvierte Logik bereitgestellt werden. Diese können die betroffenen Personen bei einem besseren Verständnis des ihnen übermittelten Scores unterstützen.

(F4) Im Vergleich zur vorhergehenden Fallgruppe F3 werden durch das System hier Handlungsempfehlungen an medizinisches Personal gegeben. Die Empfehlung des Systems kann mittelbare Auswirkungen auf die betroffene Person haben, wenn medizinisches Personal desn Handlungsempfehlungen folgt. Aus diesen Gründen sollten gegenüber Betroffenen und dem medizinischen Personal transparenzschaffende Maßnahmen zum Einsatz kommen.

(F5) Durch die starken Auswirkungen, die der Einsatz eines solchen Systems mit sich bringen kann, sollten betroffene und behandelnde Personen bestmöglich über die Funktionsweise und vorhandene Risiken unterrichtet werden. Der Einsatz transparenzschaffender Maßnahmen ist hier unumgänglich.

8.7 Ausgaben von KI-Algorithmen als personenbezogene Daten

Hintergrund

Die Ausgaben von ML-Systemen sind vielfältig und reichen von binären Ja/Nein-Entscheidungen, über Kategorieeinordnungen (bspw. die Klassifizierung eines Gegenstandes auf einem Foto), bis hin zu *Scores*, die auf beliebig großen Skalen einen bestimmten Sachverhalt einordnen. Die Kreditwürdigkeitauskunft der Schufa besteht beispielsweise aus verschiedenen Scores [SCH20], aber auch Risikoscores für bestimmte Krankheitsverläufe in der Medizin sind denkbar (siehe Fallgruppen F3 und F4 im Abschnitt 7.3).

Fragen

- In welchen Fällen kann auch das Ergebnis der Berechnung eines KI-Algorithmus als personenbezogenes Datum gewertet werden?

These

Wenn sich die Ausgaben von (KI-)Algorithmen speziell auf eine Person beziehen, sind sie genauso als personenbezogenes Datum zu werten wie etwa die (personenbezogenen) Eingabedaten. Ebenso können sie zu besonderen Kategorien personenbezogener Daten gemäß Art. 9 Abs. 1 DSGVO zählen, insbesondere wenn es sich um semantisch interpretierbare Daten handelt (z.B. ein feingranularer, medizinischen Krankheits-Risikoscore).

Ebenso fallen vergleichbar weniger sensible Berechnungsergebnisse in den Schutzbereich der DSGVO, sofern diese personenbezogen sind. So könnte auch eine binäre Ja/Nein-Entscheidung zur Kreditvergabe ein personenbezogenes Datum sein: Einerseits ist sie als Endergebnis Teil des vorangehenden Verarbeitungsprozesses. Andererseits wird der Wert durch die verantwortliche Stelle in der Regel gespeichert beziehungsweise digital kommuniziert, was ebenfalls als Verarbeitung zu werten ist.

Antithese

Wichtig für die Einstufung einer Algorithmenausgabe als personenbezogenes Datum ist der Kontext des Algorithmuseinsatzes: Liegt lediglich die Ausgabe vor, bspw. in Form eines Scores, der keine Rückschlüsse auf Personen zulässt, ist der Score selbst auch nicht als personenbeziehbares Datum zu interpretieren. Liegt der Score allerdings in Kombination mit der Algorithmeneingabe oder anderen Identitätsangaben vor, kann ersterer ein sensibler Wert sein, der ebenso wie andere personenbezogene Daten geschützt werden muss.

Für den Schutz einer Algorithmenausgabe eignen sich verschiedene Techniken des *Privacy-Preserving ML* (siehe Abschnitt 6). Beispielsweise kann die Berechnung selbst durch Homomorphe Verschlüsselung oder *Secure Multiparty Computation* in verschlüsselter Form erfolgen, sodass nur Berechtigte die Ausgabe des Algorithmus entschlüsseln können. Liegt die Ausgabe gemeinsam mit der Algorithmeneingabe vor, kann die Anonymisierung von Eingabedaten dazu beitragen, dass die Ausgabe keinem Individuum zugeordnet werden kann.

Fallgruppen

(F1) Die Algorithmenausgabe kann keinen Personenbezug haben, da keine personenbezogenen Daten verarbeitet werden.

(F2) Die Aussagekraft der Algorithmenausgabe von dem System in Fallgruppe F2 ist sehr begrenzt, daher ist der Personenbezug nicht gegeben.

(F3) Der Score, der durch das System in Fallgruppe F3 berechnet wird, kann in Kombination mit der Patientenakte bzw. den identifizierenden Daten einer Person als personenbezogenes Datum gewertet werden, da dieser den Gesundheitszustand der betroffenen Person beschreibt.

(F4) Analog zu Fallgruppe F3 kann der Score als personenbezogenes Datum gewertet werden.

(F5) Aus Medikamentationsempfehlungen lassen sich Aussagen über den Gesundheitszustand einer Person schließen. Entsprechende Daten könnten Rückschlüsse auf eine Person erlauben und sind bei Kombination mit weiteren Daten zu der betroffenen Person auch als sensibel im Sinne von Art. 9 der DSGVO angesehen werden.

8.8 Recht auf Vergessenwerden

Hintergrund

Betroffene haben laut Art. 17 DSGVO das Recht, dass ihre Daten von der verantwortlichen Stelle „unverzüglich“ gelöscht werden (siehe auch Erwägungsgründe 65 und 66). Da die Daten von Betroffenen in KI-Systemen möglicherweise tief in Modelle eingearbeitet sind, stellt sich die Frage nach der korrekten Umsetzung dieses Rechts.

Fragen

- Wie gestaltet sich die rechtskonforme Umsetzung des „Rechts auf Vergessenwerden“ in KI-Systemen?

These

Wenn personenbezogene Daten bereits verarbeitet wurden, indem ein Modell trainiert wurde und Betroffene anschließend ihr „Recht auf Vergessenwerden“ ausüben möchten, muss abgewogen werden, ob bereits trainierte, gespeicherte Modelle als personenbezogene Daten zu werten sind (analog zu Abschnitt 8.4). Selbst wenn dies jedoch der Fall sein sollte, bleibt abzuwägen, ob der Aufwand des Löschens und Neutrainierens eines KI-Modells ohne die beanstandeten Daten für die verantwortliche Stelle zumutbar ist.

Wenn für eine verantwortliche Stelle absehbar ist, dass ein Neutrainieren nicht mit verhältnismäßigem Aufwand möglich ist, muss dies Betroffenen im Rahmen der Informationspflichten mitgeteilt werden – bspw. im Text einer Einwilligungserklärung. In diesem Fall muss ausdrücklich darauf hingewiesen werden, dass die Datenverarbeitung im Rahmen des Modelltrainings irreversible Folgen hat, da die Daten im Modell manifestiert werden und anschließend nicht mehr gelöscht werden können.

Antithese

Ist der Aufwand des Neutrainierens für eine verantwortliche Stelle jedoch verhältnismäßig, kann dies durchaus gefordert werden. Neben der Erstellung eines neuen Modells (ohne Verwendung des Datensatzes, der gelöscht werden sollte) kann dies ggf. auch die Löschung und Neuberechnung von Inferenzen aus dem ursprünglich trainierten Modell miteinschließen. Dies dürfte allerdings nur in Ausnahmefällen eintreten, wenn der zu löschende Datensatz einen großen Einfluss auf das Modell hatte.

Ein weiterer Ansatz, der je nach technischer Umsetzbarkeit praktikabel sein kann, ist *Machine Unlearning* [CY15]. Diese Technik erlaubt es, den Einfluss einzelner Datensätze aus den Trainingsdaten zu eliminieren. Ob dies rechtlich einer Löschung gleichkommt, bleibt zu klären.

Fallgruppen

(F1) Da in dieser Fallgruppe keine personenbezogene Daten im Training zum Einsatz kommen, gibt es keine Betroffenen, die Anspruch auf ihr Recht auf Vergessenwerden äußern können.

(F2) Abhängig von der Vereinbarung mit dem Datensatz-Bereitstellers könnte ggf. ein Anspruch seitens der Betroffenen bestehen, ihr Recht auf Vergessenwerden durchzusetzen.

(F3) In dieser Fallgruppe könnte es von der technischen Umsetzung des KI-Systems abhängen, ob ein Recht auf Löschung einzelner Datensätze ein Neutrainieren rechtfertigt. Je nach Aufwand und unter Abwägung anderer Maßnahmen wie *Machine Unlearning* [CY15] könnte ein Neutrainineren und eine Neuberechnung der Scores nötig sein. Zudem muss ggf. die Frage geklärt werden, wie stark der Einfluss eines einzelnen Datensatzes auf die Systemausgaben sind bzw. ob ein einzelnes personenbezogenes Datum im Trainingsprozess tatsächlich im Modell manifestiert wird.

(F4) Das KI-System dieser Fallgruppe wurde mit denselben Daten trainiert wie das System in Fallgruppe F3, dementsprechend müssen hier dieselben Fragen geklärt werden.

(F5) Da in dieser Fallgruppe anonymisierte Daten verwendet werden, ist ein Personenbezug und damit eine Umsetzung von Betroffenenrechten nicht notwendig.

8.9 Recht auf Berichtigung

Hintergrund

Laut Art. 16 DSGVO haben Betroffene auch ein Recht auf unverzügliche Berichtigung ihrer personenbezogenen Daten, falls sie bei der verantwortlichen Stelle unrichtig vorliegen. Analog zu Abschnitt 8.8 stellt sich auch hier die Frage nach rechtskonformer Umsetzung, wenn die Daten von Betroffenen in der Trainingsphase eines ML-Systems genutzt werden. Dies ist insbesondere relevant, da Falschangaben in den Trainingsdaten zu fehlerhaften Modellen und damit auch zu fehlerhaften Schlussfolgerungen in der Inferenzphase führen können.

Fragen

- Was ist bei der Umsetzung des Rechts auf Berichtigung in KI-Systemen zu beachten? Müssen trainierte Modelle verworfen bzw. neu trainiert werden, wenn fehlerhafte Trainingsdaten vorlagen?

These

Bei KI-Systemen dürfte der Korrekturaufwand durch das Neu-Trainieren und ggf. Anpassen von Modellen oftmals sehr hoch sein, sodass die potenziellen schädlichen Auswirkungen von Falschinformationen im Modell entsprechend schwer wiegen müssen. Sofern beispielsweise nur einzelne Datensätze von hunderttausenden betroffen ist, sind die Auswirkungen auf das Training eines ML-Modells vermutlich überschaubar. Dies ist auch in der Natur maschineller Lernverfahren begründet. Die Zusammenhänge, die in einem Modell abgebildet werden, sind im Normalfall statistischer Natur, was auch den großen Datenbedarf maschineller Lernverfahren erklärt. Diese Natur steht im Gegensatz zu der Annahme, dass ein fehlerhafter Datensatz in den Trainingsdaten große Auswirkungen auf das trainierte Modell und damit die Genauigkeit einer späteren Vorhersage besitzt.

Als weitere Perspektive kann die These aus Abschnitt 8.4 herangezogen werden: So muss ggf. erst argumentiert werden, dass ein ML-Modell als Sammlung personenbezogener Daten gewertet werden kann. Wenn das Modell aus rechtlicher Sicht keine personenbezogenen Daten enthält, kann auch kein Recht auf Berichtigung eingefordert werden.

Antithese

Das Anwenden eines mit unrichtigen Daten trainierten Modells ist einerseits nicht mit Art. 5 Abs. 1 DSGVO vereinbar („Personenbezogene Daten müssen [...]“ lit. a: „[...] nach Treu und Glauben verarbeitet werden“ und lit. d: „sachlich richtig und erforderlichenfalls auf dem neuesten Stand sein“). Andererseits sollte es auch im Sinne der Verantwortlichen liegen, kein Modell zu verwenden, das mit unrichtigen Daten trainiert wurde.

Wie bereits beschrieben ist der Einfluss einzelner fehlerhafter Datensätze zu vernachlässigen. Sind hingegen fast alle Datensätze betroffen, beispielsweise durch einen systematischen Fehler in der Datenerhebung, ist die Wahrscheinlichkeit deutlich höher, dass ein resultierender KI-Algorithmus auch falsche Empfehlungen gibt. Diese Falschaussagen können sich unter Umständen auch als systematische Diskriminierung manifestieren, die auch trotz eines eventuell hohen Aufwandes nach einer Korrektur verlangt. Auch wenn dieser Fall nicht mehr direkt auf einer Wahrnehmung des Rechts auf Berichtigung beruht, so sind die Auswirkungen für die Verantwortlichen dieselben.

Insgesamt gilt es zwischen potenziell entstehenden Schäden (unter Rücksichtnahme auf deren Schwere und Eintrittswahrscheinlichkeiten) und dem Korrekturaufwand für die verantwortliche Stelle abzuwägen. Dies deckt sich mit dem Gutachten der Datenethikkommission der Bundesregierung [Dat19, S. 92].

Fallgruppen

Bezogen auf die Fallgruppen ergeben sich folgende Einschätzungen:

(F1) Da in dieser Fallgruppe keine personenbezogenen Daten für das Training verwendet wurden, ist das Recht auf Berichtigung hier nicht relevant.

(F2) Da die Trainingsdaten aus einem lizensierten Datensatz bestehen, sind Betroffenenrechte wohl höchstens gegenüber dem Anbieter des Datensatzes geltend zu machen. Da die Auswirkungen des Einsatzesdes beschriebenen KI-Systems als unkritisch einzuschätzen sind, wäre hier ein Neutrainieren des Modells bei fehlerhaften Daten kaum verpflichtend.

(F3) Sollten in diesem System systematisch fehlerhafte Daten für das Training verwendet worden sein, so ist eine Korrektur des Modells angeraten. Da lediglich allgemeine Empfehlungen gegeben werden, die im Normalfall vermutlich keine negativen Auswirkungen auf Betroffene haben, sind fehlerhaft erfasste Daten hier jedoch als weniger kritisch zu sehen.

(F4) Da im Vergleich zu vorhergehenden Fallgruppe direkte Behandlungsempfehlungen gegeben werden, ist eine Korrektur bei systematisch fehlerhaften Daten hier unumgänglich. Bei fehlerhaften Einzeldaten ist der Einfluss dieser Daten vermutlich gering und es kann ggf. auf eine Modellanpassung verzichtet werden. Eine Entscheidung ist hier im Einzelfall zu treffen.

(F5) Da in dieser Fallgruppe anonymisierte Daten verwendet werden, ist ein Personenbezug und damit eine Umsetzung von Betroffenenrechten nicht notwendig.

Literatur

[Aba+16] Martin Abadi u. a. *Deep learning with differential privacy*. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, S. 308–318.

[Alp19] Ethem Alpaydin. *Maschinelles Lernen*. Berlin, Boston: De Gruyter Oldenbourg, 2019. ISBN: 978-3-11-061789-4. DOI: https://doi.org/10.1515/9783110617894.

[Ang+16] Julie Angwin u. a. *Machine Bias*. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (besucht am 01. 04. 2021).

[Art17] Artikel-29-Datenschutzgruppe. *Working Paper 248*. 2017. URL: https://ec.europa.eu/newsroom/document.cfm?doc_id=47711 (besucht am 24. 11. 2020).

[Ate+15] Giuseppe Ateniese u. a. *Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers*. In: *International Journal of Security and Networks* 10.3 (2015), S. 137–150.

[Bac+15] Sebastian Bach u. a. *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation*. In: *PLOS ONE* 10.7 (Juli 2015), S. 1–46. DOI: 10.1371/journal.pone.0130140. URL: https://doi.org/10.1371/journal.pone.0130140.

[Ben75] Jon Louis Bentley. *Multidimensional Binary Search Trees Used for Associative Searching*. In: *Commun. ACM* 18.9 (Sep. 1975), S. 509–517. ISSN: 0001-0782. DOI: 10.1145/361002.361007. URL: https://doi.org/10.1145/361002.361007.

[BG18] Joy Buolamwini und Timnit Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Hrsg. von Sorelle A. Friedler und Christo Wilson. Bd. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 2018, S. 77–91.

[BK18] Andrew L Beam und Isaac S Kohane. *Big data and machine learning in health care*. In: *Jama* 319.13 (2018), S. 1317–1318.

[Bon+17] Keith Bonawitz u. a. *Practical secure aggregation for privacy-preserving machine learning*. In: *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, S. 1175–1191.

[Bra+20] Lennart Braun u. a. *MOTION-A Framework for Mixed-Protocol Multi-Party Computation.* In: *IACR Cryptol. ePrint Arch.* 2020 (2020), S. 1137.

[Bre+84] Leo Breiman u. a. *Classification and regression trees*. CRC press, 1984.

[CD+15] Ronald Cramer, Ivan Bjerre Damgård u. a. *Secure multiparty computation*. Cambridge University Press, 2015.

[CG18] Sam Corbett-Davies und Sharad Goel. *The measure and mismeasure of fairness: A critical review of fair machine learning*. In: *arXiv preprint arXiv:1808.00023* (2018).

[Cho+21] Christopher A Choquette-Choo u. a. *Label-only membership inference attacks*. In: *International Conference on Machine Learning*. PMLR. 2021, S. 1964–1974.

[CP21] José Cabrero-Holgueras und Sergio Pastrana. *SoK: Privacy-Preserving Computation Techniques for Deep Learning*. In: *Proceedings on Privacy Enhancing Technologies* 2021.4 (2021), S. 139–162.

[CV95] Corinna Cortes und Vladimir Vapnik. *Support-vector networks*. In: *Machine learning* 20.3 (1995), S. 273–297.

[CY15] Yinzhi Cao und Junfeng Yang. *Towards making systems forget with machine unlearning*. In: *2015 IEEE Symposium on Security and Privacy*. IEEE. 2015, S. 463–480.

[Dat19] Datenethikkommission der Bundesregierung. *Gutachten der Datenethikkommission*. 2019. URL: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_DE.pdf (besucht am 25. 05. 2021).

[Der19] Der Landesbeauftragte für Datenschutz und Informationsfreiheit Mecklenburg-Vorpommern. *Standard-Datenschutzmodell 2.0*. 2019. URL: https://www.datenschutz-mv.de/datenschutz/datenschutzmodell/ (besucht am 23. 12. 2020).

[DK17] Finale Doshi-Velez und Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[DLR06] D. J. DeWitt, K. LeFevre und R. Ramakrishnan. *Mondrian Multidimensional K-Anonymity*. In: *22nd International Conference on Data Engineering*. Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2006, S. 25. DOI: 10.1109/ICDE.2006.101.

[Dos+17] Finale Doshi-Velez u. a. *Accountability of AI under the law: The role of explanation*. In: *arXiv preprint arXiv:1711.01134* (2017).

[Dru+96] Harris Drucker u. a. *Support Vector Regression Machines*. In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. NIPS'96. Denver, Colorado: MIT Press, 1996, S. 155–161.

[Du+17] Zhao-Hui Du u. a. *Secure encrypted virtualization is unsecure*. In: *arXiv preprint arXiv:1712.05090* (2017).

[Duk19] Duke-Margolis Center for Health Policy. *Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care*. 2019. URL: https://healthpolicy.duke.edu/sites/default/files/2019-11/dukemargolisaienableddxss.pdf (besucht am 08. 04. 2021).

[Dwo+06] Cynthia Dwork u. a. *Calibrating noise to sensitivity in private data analysis*. In: *Theory of cryptography conference*. Springer. 2006, S. 265–284.

[DY14] Li Deng und Dong Yu. *Deep learning: methods and applications*. In: *Foundations and trends in signal processing* 7.3–4 (2014), S. 197–387.

[Eur19] Europäische Kommission. *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*. 2019. URL: https://ec.europa.eu/docsroom/documents/37581 (besucht am 07. 07. 2021).

[Fan21] Alexander Fanta. *EU verbietet automatisierte Gesichtserkennung an öffentlichen Orten – „mit wenigen Ausnahmen"*. 2021. URL: https://netzpolitik.org/2021/kuenstliche-intelligenz-eu-verbietet-automatisierte-gesichtserkennung-an-oeffentlichen-orten-mit-wenigen-ausnahmen/ (besucht am 06. 05. 2021).

[FJR15] Matt Fredrikson, Somesh Jha und Thomas Ristenpart. *Model inversion attacks that exploit confidence information and basic countermeasures*. In: *22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015, S. 1322–1333.

[FLS20] Agata Foryciarz, Daniel Leufer und Katarzyna Szymielewicz. *Black-Boxed Politics: Opacity is a Choice in AI Systems*. 2020. URL: https://en.panoptykon.org/articles/black-boxed-politics-opacity-choice-ai-systems (besucht am 29. 10. 2020).

[Gan+18] Karan Ganju u. a. *Property inference attacks on fully connected neural networks using permutation invariant representations*. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, S. 619–633.

[Geb+18] Timnit Gebru u. a. *Datasheets for datasets*. In: *arXiv preprint arXiv:1803.09010* (2018).

[Gen09] Craig Gentry. *Fully homomorphic encryption using ideal lattices*. In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009, S. 169–178.

[Gia+18] Milena A Gianfrancesco u. a. *Potential biases in machine learning algorithms using electronic health record data*. In: *JAMA internal medicine* 178.11 (2018), S. 1544–1547.

[GLN12] Thore Graepel, Kristin Lauter und Michael Naehrig. *ML confidential: Machine learning on encrypted data*. In: *International Conference on Information Security and Cryptology*. Springer. 2012, S. 1–21.

[Goo+14] Ian J. Goodfellow u. a. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[GSS14] Ian J Goodfellow, Jonathon Shlens und Christian Szegedy. *Explaining and harnessing adversarial examples*. In: *arXiv preprint arXiv:1412.6572* (2014).

[Gu+18] Jiuxiang Gu u. a. *Recent advances in convolutional neural networks*. In: *Pattern Recognition* 77 (2018), S. 354–377. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.10.013. URL: http://www.sciencedirect.com/science/article/pii/S0031320317304120.

[Hac18] Philipp Hacker. *Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law*. In: (2018).

[HAP17] Briland Hitaj, Giuseppe Ateniese und Fernando Perez-Cruz. *Deep models under the GAN: information leakage from collaborative deep learning*. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, S. 603–618.

[Har+18] Andrew Hard u. a. *Federated learning for mobile keyboard prediction*. In: *arXiv preprint arXiv:1811.03604* (2018).

[HK70] Arthur E. Hoerl und Robert W. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. In: *Technometrics* 12.1 (1970), S. 55–67.

[HN18] Thomas Hoeren und Maurice Niehoff. *KI und Datenschutz–Begründungserfordernisse automatisierter Entscheidungen*. In: *RW Rechtswissenschaft* 9.1 (2018), S. 47–66.

[Hol+17] Andreas Holzinger u. a. *What do we need to build explainable AI systems for the medical domain?* 2017. arXiv: 1712.09923 [cs.AI].

[Hol18] Andreas Holzinger. *Explainable ai (ex-ai)*. In: *Informatik-Spektrum* 41.2 (2018), S. 138–143.

[HTG16] Ehsan Hesamifard, Hassan Takabi und Mehdi Ghasemi. *Cryptodl: towards deep learning over encrypted data*. In: *Annual Computer Security Applications Conference (ACSAC 2016), Los Angeles, California, USA*. Bd. 11. 2016.

[IT20] ICO und The Alan Turing Institute. *Consultation on Explaining AI decisions guidance*. 2020. URL: https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/ (besucht am 30. 10. 2020).

[JX09] Pawel Jurczyk und Li Xiong. *Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers*. In: *Data and Applications Security XXIII*. Hrsg. von Ehud Gudes und Jaideep Vaidya. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, S. 191–207. ISBN: 978-3-642-03007-9.

[Kai+20] Georgios A Kaissis u. a. *Secure, privacy-preserving and federated machine learning in medical imaging*. In: *Nature Machine Intelligence* 2.6 (2020), S. 305–311.

[Kel20] Marcel Keller. *MP-SPDZ: A versatile framework for multi-party computation*. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 2020, S. 1575–1590.

[Knu21] Tobias Knuth. *Lernende Entscheidungsbäume*. In: *Informatik Spektrum* 44.5 (2021), S. 364–369. ISSN: 1432-122X. DOI: 10.1007/s00287-021-01398-0. URL: https://doi.org/10.1007/s00287-021-01398-0.

[Kom21] Europäische Kommission. *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. 2021. URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence (besucht am 06. 05. 2021).

[Kon16] Kongress der Vereinigten Staaten. *21st Centry Cures Act*. 2016. URL: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf (besucht am 06. 05. 2021).

[Kra+21] Tom Kraus u. a. *Erklärbare KI: Anforderungen, Anwendungsfälle und Lösungen*. 2021. URL: https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.html (besucht am 27. 05. 2021).

[Lap+19] Sebastian Lapuschkin u. a. *Unmasking clever hans predictors and assessing what machines really learn*. In: *Nature communications* 10.1 (2019), S. 1–8.

[LEL19] Scott M. Lundberg, Gabriel G. Erion und Su-In Lee. *Consistent Individualized Feature Attribution for Tree Ensembles*. 2019. arXiv: 1802.03888 [cs.LG].

[Li+20] Tian Li u. a. *Federated learning: Challenges, methods, and future directions*. In: *IEEE Signal Processing Magazine* 37.3 (2020), S. 50–60.

[Liu+17] Jian Liu u. a. *Oblivious neural network predictions via minionn transformations*. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, S. 619–631.

[LYY20] Lingjuan Lyu, Han Yu und Qiang Yang. *Threats to federated learning: A survey*. In: *arXiv preprint arXiv:2003.02133* (2020).

[Mac+06] A. Machanavajjhala u. a. *L-diversity: privacy beyond k-anonymity*. In: *22nd International Conference on Data Engineering (ICDE'06)*. Apr. 2006, S. 24–24. DOI: 10.1109/ICDE.2006.1.

[Mac+67] James MacQueen u. a. *Some methods for classification and analysis of multivariate observations*. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Bd. 1. 14. Oakland, CA, USA. 1967, S. 281–297.

[McM+17] Brendan McMahan u. a. *Communication-efficient learning of deep networks from decentralized data*. In: *Artificial Intelligence and Statistics*. PMLR. 2017, S. 1273–1282.

[Meh+21] Ninareh Mehrabi u. a. *A Survey on Bias and Fairness in Machine Learning*. In: *ACM Comput. Surv.* 54.6 (2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

[Mel+19] Luca Melis u. a. *Exploiting unintended feature leakage in collaborative learning*. In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, S. 691–706.

[MIE17] Ahmad Moghimi, Gorka Irazoqui und Thomas Eisenbarth. *Cachezoom: How SGX amplifies the power of cache attacks*. In: *International Conference on Cryptographic Hardware and Embedded Systems*. Springer. 2017, S. 69–90.

[Mit+19] Margaret Mitchell u. a. *Model cards for model reporting*. In: *Conference on fairness, accountability, and transparency*. 2019, S. 220–229.

[Mol21] Christoph Molnar. *Interpretable machine learning*. 2021. URL: https://christophm.github.io/interpretable-ml-book/ (besucht am 15. 03. 2021).

[MZ17] Payman Mohassel und Yupeng Zhang. *Secureml: A system for scalable privacy-preserving machine learning*. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, S. 19–38.

[NK19] Kee Yuan Ngiam und Wei Khor. *Big data and machine learning algorithms for health-care delivery*. In: *The Lancet Oncology* 20.5 (2019), e262–e273.

[NS08] Arvind Narayanan und Vitaly Shmatikov. *Robust De-anonymization of Large Sparse Datasets*. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. Mai 2008, S. 111–125. DOI: 10.1109/SP.2008.33.

[Ohr+16] Olga Ohrimenko u. a. *Oblivious multi-party machine learning on trusted processors*. In: *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 2016, S. 619–636.

[Pap+17] Nicolas Papernot u. a. *Practical black-box attacks against machine learning*. In: *ACM Asia conference on computer and communications security*. 2017, S. 506–519.

[Pih+18] Vasyl Pihur u. a. *Differentially-private"draw and discard"machine learning*. In: *arXiv preprint arXiv:1807.04369* (2018).

[PSA18] Trishan Panch, Peter Szolovits und Rifat Atun. *Artificial intelligence, machine learning and health systems*. In: *Journal of global health* 8.2 (2018).

[RAD+78] Ronald L Rivest, Len Adleman, Michael L Dertouzos u. a. *On data banks and privacy homomorphisms*. In: *Foundations of secure computation* 4.11 (1978), S. 169–180.

[Ram+15] Bharath Ramsundar u. a. *Massively multitask networks for drug discovery*. In: *arXiv preprint arXiv:1502.02072* (2015).

[Rav+16] Daniele Ravì u. a. *Deep learning for health informatics*. In: *IEEE journal of biomedical and health informatics* 21.1 (2016), S. 4–21.

[RHW86] David E Rumelhart, Geoffrey E Hinton und Ronald J Williams. *Learning representations by back-propagating errors*. In: *nature* 323.6088 (1986), S. 533–536.

[Ris01] Irina Rish. *An empirical study of the naive Bayes classifier*. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Bd. 3. 22. 2001, S. 41–46.

[RSG16] Marco Tulio Ribeiro, Sameer Singh und Carlos Guestrin. *Why should I trust you? Explaining the predictions of any classifier*. In: *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, S. 1135–1144.

[SAB15] Mohamed Sabt, Mohammed Achemlal und Abdelmadjid Bouabdallah. *Trusted execution environment: what it is, and what it is not*. In: *2015 IEEE Trustcom/BigDataSE/ISPA*. Bd. 1. IEEE. 2015, S. 57–64.

[Sal+18] Pedro Saleiro u. a. *Aequitas: A bias and fairness audit toolkit*. In: *arXiv preprint arXiv:1811.05577* (2018).

[Sch15] Jürgen Schmidhuber. *Deep learning in neural networks: An overview*. In: *Neural Networks* 61 (2015), S. 85–117. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2014.09.003. URL: http://www.sciencedirect.com/science/article/pii/S0893608014002135.

[SCH20] SCHUFA Holding AG. *Daten & Scoring*. 2020. URL: https://www.schufa.de/ueber-uns/daten-scoring/ (besucht am 23. 12. 2020).

[Sel+17] Ramprasaath R Selvaraju u. a. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. In: *IEEE international conference on computer vision*. 2017, S. 618–626.

[Sho+17] Reza Shokri u. a. *Membership inference attacks against machine learning models*. In: *IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, S. 3–18.

[Swe02] Latanya Sweeney. *k-anonymity: A model for protecting privacy*. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), S. 557–570.

[Tib96] Robert Tibshirani. *Regression Shrinkage and Selection Via the Lasso*. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), S. 267–288.

[You+19] Zhonghui You u. a. *Adversarial noise layer: Regularize neural network by adding noise*. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, S. 909–913.

[Zha+20] Yuheng Zhang u. a. *The secret revealer: Generative model-inversion attacks against deep neural networks*. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, S. 253–261.

Rechtsgutachten

Die Auswirkungen der Nutzung von KI-Software auf die ärztliche Haftung

von

Dr. Jan-Robert Schmidt

Inhaltsverzeichnis

A. Einführung ... 1
I. Vorwort ... 1
II. Fragestellungen und Zielsetzung ... 2
III. Begriffsbestimmungen ... 3
1. Algorithmen ... 3
2. Künstliche Intelligenz (KI) ... 3
3. Intelligenzrisiko ... 4
B. Grundsätze der Arzthaftung ... 6
I. Haftung aus Behandlungsvertrag gemäß §§ 630a, 280 Abs. 1 BGB ... 6
II. Deliktische Haftung ... 7
C. Arzthaftung bei Nutzung von KI-Software ... 9
I. Besonderheiten bei der Haftung aus Behandlungsvertrag bei Nutzung von KI-Software ... 9
1. Pflichtverletzung durch Nutzung vom KI-Software ... 9
2. Die Nutzung von selbstlernenden Algorithmen und der „ärztliche Standard“ ... 10
a. KI-Systeme als *Neulandmethoden* ... 10
b. Der Einfluss von KI-Software auf den „ärztlichen Standard“ ... 12
c. Fazit ... 14
3. Anforderungen an die ärztliche Sorgfalt bei der Nutzung von KI-Software ... 15
a. Anforderungen an den Arzt ... 15
b. Zurechnung von Maschinenfehlern über § 278 BGB ... 16
4. Fazit ... 18
II. Besonderheiten bei der deliktischen Haftung bei Nutzung vom KI-Software ... 18
1. Haftung nach § 823 BGB ... 18
2. Haftung nach § 831 BGB für KI-Software als digitalem Verrichtungsgehilfen ... 19
3. Analoge Anwendung der Tierhalterhaftung § 833 BGB ... 20
4. Analogien zu § 832 und § 836 BGB ... 21
5. Fazit ... 21
III. Der Einfluss der Nutzung von KI-Software auf die Aufklärungspflicht nach § 630e BGB ... 21
IV. Der Einfluss der Nutzung von KI-Software auf die Beweislast nach § 630h BGB ... 23
V. Perspektiven auf die zukünftige Rechtsentwicklung ... 24
1. Vorschläge aus der Literatur ... 24
a. E-Person ... 24
b. Fondlösung ... 25
2. Reformvorhaben der EU-Kommission ... 25
D. Zusammenfassung der Ergebnisse ... 28

Literaturverzeichnis

Armbrüster, Christian/Prill, Jonathan: Einsatz von KI im Versicherungssektor – mit Schwerpunkt Versicherungsmedizin, ZVersWiss 2022, 177.

Baumgärtel, Gottfried/Laumen, Hans-Willi/Prütting, Hanns (Hrsg.): Handbuch der Beweislast Band 2 §§ 1-811 BGB, Carl Heymanns, 5. Auflage 2023.

Bomhard, David/Siglmüller, Jonas: Europäische KI-Haftungsrichtlinie, RDi 2022, 506.

Bördner, Jonas: Digitalisierung im Gesundheitswesen – eine haftungsrechtliche Bestandsaufnahme, GuP 2019, 131.

Borges, Georg: Rechtliche Rahmenbedingungen für autonome Systeme, NJW 2018, 977, 981.

Brand, Oliver: Haftung und Versicherung beim Einsatz von Robotik und Medizin in der Pflege, MedR 2019, 943.

Burchardi, Sophie: Risikotragung für KI-Systeme, EuZW 2022, 685.

Chibanguza, Kuuya/Kuß, Christian/Steege, Hans (Hrsg.): Künstliche Intelligenz – Recht und Praxis automatisierter und autonomer Systeme, Nomos, 2022.

Dettling, Heinz-Uwe: Künstliche Intelligenz und digitale Unterstützung ärztlicher Entscheidungen in Diagnostik und Therapie, PharmR 2019, 633.

Erdmann, Pia/Fischer, Tobias/Raths, Susan/Flessa, Steffen: Medizinische Versorgung: Systemmedizin – Herausforderungen eines aktuellen Ansatzes, Deutsches Ärzteblatt 2015, A1330.

Ernst, Anna Maria: Rechtsfragen der Systemmedizin, Springer, 2020.

Europäische Kommission (Hrsg.): Weissbuch zur künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen, COM (2020) 65, 2020.

Europäische Kommission: Richtlinienvorschlag zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz, 2022/0303/COD, 2022.

Europäisches Parlament: Zivilrechtliche Regelungen im Bereich Robotik, 2018/C252/25, 2017.

Europäisches Parlament: Entschließung des Europäischen Parlaments vom 20. Oktober 2020 mit Empfehlungen an die Kommission für eine Regelung der zivilrechtlichen Haftung beim Einsatz künstlicher Intelligenz, 2020/2014(INL), 2020.

Funer, Florian: An den Grenzen (il)legitimier Diskriminierung durch algorithmische Entscheidungsprozesse in der Medizin, in: Loh/Grote (Hrsg.): Medizin – Technik – Ethik, J.B. Metzler, 2023, S. 59-86.

Friele, Jannes/Jannes, *Christiane/Woopen, Christiane*: Algorithmen in der digitalen Gesundheitsversorgung – Eine interdisziplinäre Analyse, Bertelsmann Stiftung, 2018.

Frost, Yannick/Kießling, Marlene: Künstliche Intelligenz im Bereich des Gesundheitswesens und damit verbundene haftungsrechtliche Herausforderungen, MPR 2020, 178.

Hahn, Erik: Das „Recht auf Nichtwissen“ des Patienten bei algorithmengesteuerter Auswertung von Big Data, MedR 2019, 197.

Hart, Dieter: Ärztliche Leitlinien – Definitionen, Funktionen, rechtliche Bewertungen, MedR 1998, 8.

Hart, Dieter: Haftungsrecht und Standardbildung in der modernen Medizin – e:med und Probleme der Definition des Standards, MedR 2016, 669.

Hau,Wolfgang/Poseck Roman: BeckOK BGB, C.H.Beck, 65. Edition 2022.

Helle, Katrin: Intelligente Medizinprodukte: Ist der geltende Rechtsrahmen noch aktuell?, MedR 2020, 993.

Jansen, Christoph: Der Medizinische Standard, Springer, 2019.

Jörg, Johannes: Digitalisierung in der Medizin – Wie Gesundheits-Apps, Telemedizin, künstliche Intelligenz und Robotik das Gesundheitswesen revolutionieren, Springer, 2018.

Katzenmeier, Christian: Big Data, E-Health, M-Health, KI und Robotik in der Medizin – Digitalisierung des Gesundheitswesens – Herausforderung des Rechts, MedR 2019, 259.

Katzenmeier, Christian: KI in der Medizin – Haftungsfragen, MedR 2021, 859.

Keil, Miriam: Rechtsfragen der individualisierten Medizin, Springer, 2015.

Kienzle, Hans-Friedrich: Standards in der Medizin – Sicht des medizinischen Sachverständigen, in: Jansen/Katzenmeier/Woopen (Hrsg.), Medizin und Standard, Springer, 2020, S. 37-44.

Kniepert, Cornelius: Befunderhebung oder Diagnose – Zur Abgrenzung des Befunderhebungsfehlers vom Diagnosefehler und deren Auswirkung auf die ärztliche Praxis, Nomos, 2020.

Mühlböck, Luisa/Taupitz, Jochen: Haftung für Schäden durch KI in der Medizin, AcP 221 (2021), 179.

Münchener Kommentar zum BGB, C.H. Beck, 9. Auflage 2023.

Plagemann, Hermann: Kritik der medizinischen Vernunft – Medizinphilosophie-Biopolitik-Recht, GuP 2019, 96.

Pomberger, Gustav/Dobler, Heinz: Algorithmen und Datenstrukturen – eine systematische Einführung in die Programmierung, Pearson, 2008.

Rasche, Christoph/Reinecke, Adriana/Margaria, Tiziana: Künstliche Intelligenz im Gesundheitswesen als Kernkompetenz? Status quo, Entwicklungslinien und disruptives Potenzial, in: Pfannstiel (Hrsg.), Künstliche Intelligenz im Gesundheitswesen, Springer, 2022, S. 49-80.

Riehm, Thomas/Meier, Stanislaus: Künstliche Intelligenz im Zivilrecht, in Fischer/Hoppen/Wimmers: DGRI Jahrbuch 2018, Otto Schmidt, 2018, S. 1-36.

Spindler, Gerald: Medizin und IT, insbesondere Arzthaftungs- und IT-Sicherheitsrecht, in Katzenmeier (Hrsg.): Festschrift für Dieter Hart, Springer, 2020, S. 581-602.

Staudinger BGB, Otto Schmidt, Neubearbeitung 2021.

Ströbel, Lukas/ Grau, Robert: KI-gestützte Medizin-Apps - Rechtliche Herausforderungen eines interdisziplinären Produkts, ZD 2022, 599.

Taupitz, Jochen: Medizinische Informationstechnologie, leitliniengerechte Medizin und Haftung des Arztes, AcP 211 (2011), 352.

Teubner, Gunther: Digitale Rechtssubjekte? Zum privatrechtlichen Status autonomer Softwareagenten, AcP 218 (2018), 155.

Unabhängige Hochrangige Expertengruppe für künstliche Intelligenz: Eine Definition der KI – Wichtigste Fähigkeiten und Wissenschaftsgebiete, 2022, https://www.bundesnetzagentur.de/DE/Fachthemen/Digitalisierung/Mittelstand/Downloads/Experten.html (zuletzt geprüft: 23.02.2023).

Wagner, Gerhard: Produkthaftung für autonome Systeme, AcP 217 (2017), 707.

Wagner, Gerhard: Verantwortlichkeit im Zeichen digitaler Techniken, VersR 2020, 717.

Wagner, Gerhard: Die Richtlinie über KI-Haftung: Viel Rauch, wenig Feuer, JZ 2023, 123.

Woopen, Christiane: Medizin und Standard – Ethische Überlegungen, in: Jansen/Katzenmeier/Woopen (Hrsg.): Medizin und Standard – Verwerfungen und Perspektiven, Springer, 2020, S. 119-130.

Zech, Herbert: Entscheidungen digitaler autonomer Systeme: Empfehlen sich Regelungen zu Verantwortung und Haftung, Gutachten zum 73. DJT 2020, C.H. Beck, 2020.

Zech, Herbert: Künstliche Intelligenz und Haftungsfragen, ZfPW 2019, 198.

Zech, Herbert: Zivilrechtliche Haftung für den Einsatz von Robotern – Zuweisung von Automatisierungs- und Autonomierisiken, in: Gless/Seelmann, Intelligente Agenten und das Recht, Nomos, 2017, S. 163-204.

Zech, Herbert/Hünefeld, Isabelle Céline: Einsatz von KI in der Medizin: Haftung und Versicherung, MedR 2023, 1.

A. Einführung

I. Vorwort

Seit einigen Jahren sind „Big Data“ und die Nutzung von Algorithmen zunehmend in den öffentlichen Fokus gerückt. Der Informationsgewinnung durch Verarbeitung großer Datenmengen kommt heute in vielen Lebensbereichen wie etwa dem autonomen Fahren eine erhebliche Bedeutung zu. Auch in der Medizin wurden die Potenziale der Nutzung großer Mengen an Patientendaten für Diagnostik und Therapie erkannt. Vor dem Hintergrund neuer Herausforderungen im Behandlungsalltag gerade aufgrund knapper Personalressourcen,[1] kann der Einsatz von künstlicher Intelligenz (KI) und Algorithmen dazu beitragen, Prozesse zu beschleunigen und das medizinische Personal zu entlasten.[2] Gleichzeitig bestehen durch die Nutzung von Algorithmen auch gänzlich neue Möglichkeiten, heute schon vorhandene Daten für Diagnostik und Therapie nutzbar zu machen.[3] So können Algorithmen beispielsweise durch die Analyse einer großen Menge an Vergleichswerten zu individuell auf den Patienten abgestimmten Behandlungsvorschlägen führen.[4] In der Nutzung von algorithmenbasierten Tools und anderen digitalen Gesundheitsanwendungen wird daher auch ein wesentlicher Baustein für die Medizin der Zukunft gesehen.[5] Dies hat auch der Gesetzgeber erkannt und im Jahre 2019 durch die Schaffung von § 33a SGB V einen Rechtsrahmen für die Erstattungsfähigkeit von digitalen Medizinprodukten geschaffen, was die Entwicklungen weiterer Anwendungen in diesem Bereich begünstigen wird.[6] Aber schon heute gehört die Anwendung von künstlicher Intelligenz teilweise zum Klinikalltag.[7]

Die Nutzung von algorithmenbasierten Anwendungen wirft jedoch einige rechtliche Fragen auf, von denen teilweise unklar ist, ob sie durch die bestehenden Normen zufriedenstellend beantwortet werden können. Dies gilt insbesondere für das Haftungsrecht.[8] Bei Schädigungen eines Patienten im Rahmen einer medizinischen Behandlung stellt sich häufig die Frage, ob der behandelnde Arzt nicht einen Fehler bei der Therapie gemacht hat, die den Schaden verursacht hat und für den er zu haften hat. Diese Haftung stellt die Kehrseite der grundsätzlich bestehenden Therapiefreiheit des Arztes dar und begrenzt diese. Unklar ist allerdings, welche

---

[1] *Brand*, MedR 2019, 943, 944.
[2] *Ströbel/Grau*, ZD 2022, 599, 600.
[3] *Jörg*, Digitalisierung in der Medizin, S 89f.
[4] Ebd.
[5] *Friele/Jannes/Woopen*, Algorithmen, S. 82, siehe zum Einsatz von KI generell: *EU-Kommission*, Weißbuch, COM (2020) 65 final;
[6] s. hierzu: *Ströbel/Grau*, ZD 2022, 599, 600; *Bördner*, GuP 2019, 131ff.
[7] *Ströbel/Grau*, 599, 600.
[8] *Europäisches Parlament*, Entschließung zu zivilrechtlichen Regelungen im Bereich Robotik, Erwägungen Y und Z.

Folgen es für die ärztliche Haftung hat, wenn nicht dieser selbst die Behandlungsentscheidung trifft, sondern diese maßgeblich durch eine KI-basierte Software getroffen wird, die durch Auswertung einer großen Zahl von Daten gezielt auf den individuellen Patienten zugeschnittene Therapievorschläge macht. Aufgrund der hohen Anzahl der ausgewerteten Daten kann ein fehlerfrei programmierter Algorithmus eine höhere Entscheidungsevidenz für sich beanspruchen, als ein einzelner Mediziner, dessen Entscheidungen auf erworbener Erfahrung und Wissen aus Lektüre von Fachliteratur fußen. Allerdings ist es für den einzelnen Nutzer in aller Regel nicht erkennbar, wie ein Algorithmus zu seinem Ergebnis gekommen ist (Blackbox-Effekt).[9] Dies gilt insbesondere für KI-basierte, selbstlernende Algorithmen.[10] Es stellt sich insofern die Frage, wie es um die Therapiefreiheit eines Arztes bestellt ist, dem man einen solchen Algorithmus an die Seite stellt.[11] Zudem ist zu klären, ob das derzeitige Haftungsregime des Arzthaftungsrechts die Verantwortungsverteilung in einer solchen Behandlungssituation noch korrekt abbildet und welche Haftungsfolgen sich aus dem Einsatz einer algorithmenbasierten Software in der täglichen Arbeit ergeben.

II. Fragestellungen und Zielsetzung

Das vorliegende Rechtsgutachten befasst sich mit den arzthaftungsrechtlichen Aspekten bei der Verwendung von algorithmenbasierten Diagnose- und Therapietools. Hierfür sind zunächst die Begrifflichkeiten KI und Algorithmus, Blackbox-Effekt und Intelligenzrisiko zu klären. Im Anschluss wird die grundsätzliche ärztliche Haftung für Patientenschäden kurz skizziert. Darauf aufbauend wird analysiert, welche Folgen die Verwendung von algorithmenbasierten Diagnose- und Therapietools nach dem geltenden Recht hat. Hierbei wird ein besonderer Fokus auf die Frage gelegt, welche Auswirkungen algorithmenbasierte Tools auf den „ärztlichen Standard“ haben können. Weiterhin wird auch die Frage analysiert, ob die Verwendung von algorithmenbasierten Tools Einfluss auf die Anwendung der Beweislastregeln des Arzthaftungsrechts haben kann. Zudem wird auch der Einfluss der Nutzung solcher Tools auf die ärztliche Aufklärungspflicht nach § 630e BGB thematisiert. Zuletzt wird ein Ausblick auf die mögliche zukünftige Rechtsentwicklung gegeben.

Da es im vorliegenden Gutachten um die ärztliche Haftung geht, werden keine Ausführungen zu Besonderheiten der Haftung anderer Personen bei der Nutzung algorithmenbaiserter

[9] Siehe A. III. 2.
[10] *Katzenmeier*, MedR 2019, 259, 269.
[11] *Ernst*, Rechtsfragen der Systemmedizin, S. 145.

Diagnose- und Therapietools gemacht. Dies betrifft in erster Linie die Haftung des Herstellers solcher Softwarelösungen.[12]

III. Begriffsbestimmungen

Vor der eingehenden Betrachtung der Haftungsfragen, die den Schwerpunkt dieses Gutachtens bilden, ist es notwendig zunächst die verwendeten Begrifflichkeiten zu klären.

1. Algorithmen

Bei Algorithmen handelt es sich grob gesagt um Handlungsvorschriften zur Lösung eines spezifischen Problems. Eine etwas differenziertere Definition spricht von Algorithmen als:

„eine vollständige, präzise und in einer Notation oder Sprache mit exakter Definition abgefasste, endliche Beschreibung eines schrittweisen Problemlösungsverfahrens zur Ermittlung gesuchter Datenobjekte (ihrer Werte) aus gegebenen Werten von Datenobjekten, in dem jeder Schritt aus einer Anzahl ausführbarer, eindeutiger Aktionen und einer Angabe über den nächsten Schritt besteht.“[13]

Ein Algorithmus stellt damit je nach Komplexität letztlich eine Ansammlung von Wenn-dann-Vorgaben dar, mittels derer durch die Auswertung von Daten Ergebnisse erzeugt werden.

Eine Legaldefinition gibt es jedoch nicht. Für das vorliegende Gutachten wird von der gerade zitierten Definition ausgegangen, wenn von Algorithmen die Rede ist.

2. Künstliche Intelligenz (KI)

Ebenso wenig wie für den Begriff „Algorithmus“ gibt es auch für den Begriff „Künstliche Intelligenz“ eine einheitliche Definition.[14] Im vorliegenden Gutachten wird der Begriff „Künstliche Intelligenz“ (KI) entsprechend einem Vorschlag der von der EU-Kommission eingesetzten *Unabhängigen Hochrangigen Expertengruppe für Künstliche Intelligenz* wie folgt verstanden:

„Systeme der künstlichen Intelligenz (KI-Systeme) sind vom Menschen entwickelte Softwaresysteme (und gegebenenfalls auch Hardwaresysteme), die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene handeln, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren,

---

[12] Siehe hierzu unter anderem: Chibanguza/Kuß/Steege-*Chibanguza*, KI, § 4 E Rn 1ff.; *Frost/Kießling*, MPR 2020, 178ff; *Katzenmeier*, MedR 2021, 859, 863ff.
[13] *Pomberger/Dobler*, Algorithmen und Datenstrukturen, S. 33.
[14] *Ströbel/Grau*, ZD 2022, 599, 600; Dettling, PharmR 2019, 633, 634.

Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten, und über das bestmögliche Handeln zur Erreichung des vorgegebenen Ziels entscheiden.“[15]

Danach zeichnet sich künstliche Intelligenz unter anderem dadurch aus, dass sie selbstlernend ist und in der Lage ist, sich durch die Analyse von verarbeiteten Daten kontinuierlich weiterzuentwickeln.[16] KI-Software arbeitet dabei mit Algorithmen.[17] Diese unterscheiden sich allerdings von denen bei klassischer determinierter Software. Bei dieser sind die oben beschriebenen Wenn-dann-Vorgaben menschlicher Natur und basieren auf den Regeln, die die Programmierer dem System gegeben haben. Bei KI-Software geben die Programmierer den Algorithmen nur eine gewisse Struktur und Methodik vor. Das Ziehen von Wenn-Dann-Schlüssen bleibt jedoch der künstlichen Intelligenz vorbehalten.[18] Die von der KI gebildeten Entscheidungsregeln sind dabei für den Programmanwender in der Regel nicht nachvollziehbar. In diesem Zusammenhang wird daher auch häufig von KI als einer „Blackbox“ bzw. vom „Blackbox-Effekt“ gesprochen.[19]

Gleichzeitig heißt dies aber nicht, dass eine KI anhand der Datenlage zwingend unvoreingenommene Entscheidungen trifft. Vielmehr kann KI-Software menschliche Fehlschlüsse fortschreiben bzw. perpetuieren. Zwar handelt ein Algorithmus aufgrund der Datenlage objektiv, allerdings können von Menschen erhobene, fehlerhafte oder vorurteilsbelastete Daten, die in die Software eingespeist werden, dazu führen, dass die Software die darin enthaltenen menschlichen Fehlschlüsse beibehält und schlimmstenfalls hochskaliert.[20] Eine weitergehende Auseinandersetzung mit diesen Problematiken würde jedoch den Rahmen dieses Gutachtens sprengen.

Wenn im Folgenden von KI-Software, selbstlernenden Algorithmen und algorithmenbasierten Tools die Rede ist, so sind damit – im Sinne des vorher gesagten – Softwaresysteme gemeint, die in der Lage sind, aus den eingespeisten Daten selbstständige Schlüsse zu ziehen.

3. Intelligenzrisiko

Mit dem Begriff Intelligenz- bzw. Autonomierisiko ist gemeint, dass durch die selbstständige Weiterentwicklung von selbstlernenden Algorithmen, deren Entscheidungen nicht mehr in

[15] *HEG*, Eine Definition der KI, S. 6.
[16] *Frost/Kießling*, MPR 2020, 178, 179.
[17] Siehe A.III.1.
[18] *Dettling*, PharmR 2019, 633, 635; Brand, MedR 2019, 943, 947.
[19] *Dettling*, PharmR 2019, 633, 635; Katenmeier, MedR 2019, 259, 269; Funer, in: Loh/Grote: Medizin – Technik – Ethik, S. 60; Mühlböck/Taupitz, AcP 221 (2021), 179, 183.
[20] *Rasche/Reinecke/Margaria*, in: Pfannstiel, Künstliche Intelligenz im Gesundheitswesen, S. 61; *Katenmeier*, MedR 2019, 259, 269; *Funer*, in: Loh/Grote: Medizin – Technik – Ethik, S. 61.

Gänze vorhersehbar sind (Blackbox-Effekt).[21] Hierdurch besteht das Risiko, dass aufgrund der selbstständigen Handlungen eines KI-Systems Schäden entstehen, für die rechtlich niemand die Verantwortung trägt, da niemandem ein Verschulden zugerechnet werden kann.[22]

[21] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 17. Siehe A III 2.
[22] *Burchardi*, EuZW 2022, 685; MüKoBGB-*Wagner*, Vor § 630a Rn 61

B. Grundsätze der Arzthaftung

Die ärztliche Haftung für Pflichtverletzungen gegenüber Patienten kann sich aus Vertrags- und Deliktsrecht ergeben. Nachfolgend werden zunächst die Grundzüge der ärztlichen Haftung dargestellt, um im Weiteren (s. C.) auf die Besonderheiten der ärztlichen Haftung bei Nutzung von selbstlernenden Algorithmen bei der Behandlung einzugehen.

I. Haftung aus Behandlungsvertrag gemäß §§ 630a, 280 Abs. 1 BGB

Grundsätzlich schuldet der Arzt dem Patienten aus dem mit ihm geschlossenen Behandlungsvertrag Aufklärung und Behandlung nach Einwilligung.[23] Die Behandlung umfasst dabei den gesamten Verlauf ärztlichen Handelns von der Diagnose bis zur Nachsorge. [24] Pflichtverletzungen des Arztes während der Behandlung können zu einer Schadensersatzpflicht gemäß §§ 630a, 280 Abs. 1 BGB führen. Eine Pflichtverletzung liegt regelmäßig dann vor, wenn die Behandlung sorgfaltspflichtwidrig erfolgt ist.[25] Der allgemeine Sorgfaltsmaßstab bemisst sich dabei grundsätzlich nach § 276 Abs. 2 BGB, wird aber im Arzthaftungsrecht durch § 630a Abs. 2 BGB konkretisiert.[26] Hiernach hat eine Behandlung

nach den zum Zeitpunkt der Behandlung bestehenden, allgemein, anerkannten Standards zu erfolgen, soweit nichts anderes vereinbart ist.

Der ärztliche „Standard“ ist somit das, woran sich die Heilbehandlung messen lassen muss, und damit auch Maßstab bei der Prüfung einer möglichen Schadensersatzpflicht, da eine Verletzung des „Standards“ eine Pflichtverletzung indiziert.[27] Eine pauschale Definition des geschuldeten „Standards“ gibt es dabei aufgrund der stets individuellen Behandlungssituation nicht.[28] Nach ständiger Rechtsprechung des BGH wird der „Standard“ grundsätzlich wie folgt beschrieben:

„Der Standard gibt Auskunft darüber, welches Verhalten von einem gewissenhaften und aufmerksamen Arzt in der konkreten Behandlungssituation aus der berufsfachlichen Sicht seines Fachbereichs im Zeitpunkt der Behandlung erwartet werden kann. Er repräsentiert den jeweiligen Stand der naturwissenschaftlichen Erkenntnisse und der ärztlichen Erfahrung, der zur

---

[23] MüKoBGB-*Wagner*, § 630a, Rn 125ff.; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 5.
[24] MüKoBGB-*Wagner*, § 630a, Rn 127.
[25] MüKoBGB-*Wagner*, § 630a, Rn 125.
[26] BeckOKBGB-*Katzenmeier*, § 630a, Rn 145; MüKoBGB-*Wagner*, § 630a, Rn 2; Staudinger-*Gutmann*, § 630a, Rn 131,142; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 6; *Kniepert*, Befunderhebung, S. 31 FN 81 m.w.N.
[27] MüKoBGB-*Wagner*, § 630a, Rn 152; *Hart*, MedR 2016, 669, 671.
[28] Staudinger-*Gutmann*, § 630a, Rn 132.

Erreichung des ärztlichen Behandlungsziels erforderlich ist und sich in der Erprobung bewährt hat."[29]

Hiernach gibt es verschiedene Faktoren, die den „Standard" bestimmen, wie die konkrete Behandlungssituation, den Zeitpunkt der Behandlung und die berufsfachliche Sicht des Fachbereichs.[30] Im zweiten Satz der Standardbeschreibung des BGH werden zwei Erkenntnisquellen ärztlicher Arbeit benannt. Zum einen der *Stand der naturwissenschaftlichen Kenntnisse* und zum anderen die *ärztliche Erfahrung*. Für die Standardbestimmung soll es also sowohl auf objektive wissenschaftliche Erkenntnisse, als auch auf die konkrete Erfahrung des einzelnen Behandelnden ankommen.[31]

Ob eine Methode dem „ärztlichen Standard" entsprochen hat, ist im Arzthaftungsprozess dabei grundsätzlich durch einen Gutachter zu klären.[32] Hilfe bei der Einschätzung einer Behandlungsmethode und gleichzeitig eine gewisse Absicherung im Hinblick auf die Arzthaftung bilden die ärztlichen Leitlinien.[33] Bei den Leitlinien handelt es sich um meist von medizinischen Fachgesellschaften herausgegebene Entscheidungshilfen zur Vorgehensweise bei bestimmten medizinischen Problemstellungen.[34] Sie können zur Konkretisierung des Sorgfaltsmaßstabs herangezogen werden, haben aber keine Bindungswirkung für Gerichte und sind nicht mit dem „Standard" gleichzusetzen, da sie veralten und ihre Hinweise von neuen Behandlungsmethoden verdrängt werden können.[35] Um eine Pflichtverletzung zu vermeiden, ist es jedoch auch nicht notwendig, stets die von der Mehrheit angewandte Behandlungsmethode zu wählen. Auch neuartige Behandlungsmethoden können im Rahmen des § 630a Abs. 2 BGB angewandt werden. Allerdings gelten hier besondere Anforderungen an die ärztliche Sorgfalt im Hinblick auf die Risikoabwägung.[36]

II. Deliktische Haftung

Neben der Haftung aus Vertrag kommt bei ärztlichem Fehlverhalten auch eine Haftung aus Deliktsrecht gemäß § 823 Abs. 1 BGB und § 823 Abs. 2 BGB iVm § 222 bzw. § 229 StGB in Betracht.[37] Da die Arzthaftung aus Behandlungsvertrag nach §§ 280 Abs. 1, 630a Abs. 1 BGB

[29] BGH NJW 2016, 713, 714; BGHZ 102, 17 [24 f.] = NJW 1988, 763; NJW-RR 2014, 1053 = VersR 2014, 879 Rn. 11.
[30] S. hierzu vertiefend: *Jansen*, Der Medizinische Standard, S. 43ff.
[31] *Plagemann*, GuP 2019, 96, 104f.
[32] BeckOKBGB-*Katzenmeier*, § 630a, Rn 151; Staudinger-*Gutmann*, § 630a, Rn 135.
[33] *Kniepert*, Befunderhebung, S.33ff; *Dettling*, PharmR 2019, 633, 634.
[34] MüKoBGB-*Wagner*, § 630a, Rn 150; BeckOKBGB-*Katzenmeier*, § 630a, Rn 154.
[35] Staudinger-*Gutmann*, § 630a, Rn 153ff.; BeckOKBGB-*Katzenmeier*, § 630a, Rn 155.
[36] MüKoBGB-*Wagner*, § 630a, Rn 158; Staudinger-*Gutmann*, § 630a, Rn 146.
[37] BeckOKBGB-*Katzenmeier*, § 630a, Rn 11ff.; MüKoBGB-*Wagner*, § 823, Rn 1082ff.

wesentlich aus dem richterlich ausgeformten Deliktsrecht entwickelt wurde, laufen vertragliche Haftung und deliktische Haftung des Arztes zurzeit in den allermeisten Fällen parallel.[38]

[38] BeckOKBGB-*Förster*, § 823, Rn 791.; MüKoBGB-*Wagner*, § 823, Rn 1082ff.

C. Arzthaftung bei Nutzung von KI-Software

Im Zusammenhang mit der Nutzung von KI-Software mit selbstlernenden Algorithmen stellen sich verschiedene Fragen im Hinblick auf die oben skizzierten Prinzipien des Arzthaftungsrechts, wie beispielsweise, ob solche Programme überhaupt verwendet werden dürfen und in welchem Verhältnis sie zum ärztlichen „Standard“ stehen. Auch stellt sich die Frage, ob und wenn ja in welcher Form Fehlfunktionen des Programms dem verwendenden Arzt zuzurechnen sind. Bei der nachfolgenden Erörterung wird von der Konstellation der Nutzung eines Programms mit einem selbstlernenden Algorithmus durch einen Arzt im Rahmen einer Heilbehandlung ausgegangen. Bei der Verwendung eines solches Programms trägt der Arzt hierfür grundsätzlich die Verantwortung gegenüber dem Patienten.[39] Die Verwendung hat also auch direkten Einfluss auf die Arzthaftung.

I. Besonderheiten bei der Haftung aus Behandlungsvertrag bei Nutzung von KI-Software

1. Pflichtverletzung durch Nutzung vom KI-Software

Die Nutzung einer KI-Software mit einem selbstlernenden Algorithmus könnte für sich genommen schon eine Pflichtverletzung darstellen.[40] Als Argument hierfür wird vorgebracht, dass selbstlernende Algorithmen für den Menschen nicht mehr zu durchschauen seien (s. „Black-box-Effekt“ unter A.III.2.) und daher auch nicht mehr zu beherrschen seien.[41] Eine Nutzung würde daher nicht kalkulierbare Risiken mit sich bringen.

Demgegenüber hält der weit überwiegende Teil der Literatur den Einsatz von KI-Software nicht per se für pflichtwidrig.[42] Hierzu wird vorgebracht, dass Sorgfaltspflichten nicht so weit gehen würden, dass eine absolute Sicherheit gewährleistet werden müsse.[43] Geschuldet sei vielmehr die rechtsübliche Sorgfalt.[44] Dies gelte auch für das Arzthaftungsrecht, da ärztliches Handeln ansonsten in vielen Fällen gar nicht möglich wäre.[45] Bei der Nutzung von KI-Software sei daher nur zu verlangen, dass diese denselben Sicherheitsstandard biete, wie ein von Menschen

[39] *Bördner*, GuP 2019, 131, 133.
[40] *Zech*, in: Gless/Seelmann, Intelligente Agenten und das Recht, S. 163, 191ff.
[41] *Zech*, in: Gless/Seelmann, Intelligente Agenten und das Recht, S. 163, 191ff.; *Zech*, Entscheidungen digitaler autonomer Systeme, S. A 55; *Teubner*, AcP 218 (2018), 155, 185f.
[42] MüKoBGB-Wagner, Vor § 630a, Rn 60; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 Rn 12; *Wagner*, AcP 217 (2017), 707, 728; *Wagner*, VersR 2020, 717, 727; *Katzenmeier*, MedR 2021, 859, 860f.
[43] *Wagner*, AcP 217 (2017), 707, 728; *Wagner*, VersR 2020, 717, 727.
[44] S. BGH, NJW 2013, 48; BGH NJW 2014, 2104, 2105.
[45] Katzenmeier, MedR 2021, 859, 860.

gesteuertes System.[46] Dem ist zuzustimmen, da die Annahme einer prinzipiellen Pflichtwidrigkeit der Nutzung eines solchen Systems eine nicht bestehende Gefährdungshaftung herbeikonstruieren würde und gegen das Verschuldensprinzip verstoßen würde.[47]

Dass die Nutzung von künstlicher Intelligenz nicht als grundsätzlich pflichtwidrig einzustufen ist, entbindet den Verwendenden jedoch nicht von der Verpflichtung, stets eine Abwägung zwischen den Vorteilen, die der Einsatz mit sich bringen kann und den Risiken treffen zu müssen.[48] Der Verwender ist gleichwohl – wie bei anderen medizinischen Geräten auch – verpflichtet, sich mit der generellen Funktionsweise der KI-Software vertraut zu machen.[49] Gleichsam trifft den Verwender eine Instandhaltungs- und Wartungspflicht, auf die vor dem Hintergrund des „Intelligenzrisikos[50]“ der KI ein besonderes Augenmerk zu legen ist.[51]

2. Die Nutzung von selbstlernenden Algorithmen und der „ärztliche Standard“

a. KI-Systeme als *Neulandmethoden*

Wie gerade festgestellt, ist die Nutzung von KI-Programmen noch nicht grundsätzlich pflichtwidrig.[52] Pflichtwidrigkeit kann jedoch aus der Nichteinhaltung des ärztlichen „Standards“ folgen.[53] Die Bildung des „Standards“ befindet sich dabei in einem Spannungsverhältnis zwischen der Orientierung an Bewährtem und Innovation.[54] Die Medizin entwickelt ihre Methoden stetig weiter. Ein statischer „Standard“, der nicht an aktuelle Entwicklungen angepasst wird, wäre daher immer nur eine (schnell veraltende) Momentaufnahme. Vor diesem Hintergrund sind auch Methoden, die noch neu sind und noch keine breite Anwendung gefunden haben (*Neulandmethoden*), nicht unzulässig.[55]

Die behandelnde Person hat jedoch eine sorgfältige Abwägung zwischen den Methoden, die zum aktuellen „Standard“ gehören und der neuen Methode durchzuführen, die zu dem Ergebnis kommen muss, dass der Nutzen der neuen Methode etwaige Risiken überwiegt.[56] Bei einem KI-Programm, das durch Auswertung großer Datenmengen auf den individuellen

---

[46] *Wagner*, VersR 2020, 717, 727.
[47] s. hierzu auch: *Katzenmeier*, MedR 2021, 859, 860f.
[48] *Riehm/Meier*, in: Fischer/Hoppen/Wimmers, DRGI Jahrbuch 2018, Rn 27.
[49] *Spindler*, in: Katzenmeier (Hrsg.), FS Hart, 581, 587.
[50] Siehe A.III.3.
[51] Ebd. S. 588
[52] Siehe C.I.1.
[53] Siehe B.I.; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 13.
[54] *Hart*, MedR 2016, 669, 671.
[55] Siehe B.I.
[56] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 13, 33; *Ernst*, Rechtsfragen der Systemmedizin, S. 143.

Patienten abgestimmte Behandlungsvorschläge macht, liegt ein konkreter Mehrwert vor. Ein solches Programm wertet (bei korrekter Funktionsweise) wesentlich mehr Daten für die Entscheidungsfindung aus, als es einem einzelnen Menschen möglich wäre.[57] Hierdurch kann die Behandlung enger auf den einzelnen Patienten abgestimmt werden, als es bei den Handlungsanweisungen des „klassischen" „Standards" der Fall ist.[58] Dabei wird man davon auszugehen können, dass eine Behandlung, die das Individuum mehr in den Blick nimmt, für dieses auch risikoärmer ist. Dies wäre ein starkes Argument für die Nutzung von KI-Software als *Neulandmethode* im Rahmen der oben skizzierten Abwägungsentscheidung.

Dies gilt jedoch nur, wenn der Verwender davon ausgehen kann, dass die KI-Software auch tatsächlich richtig funktioniert, das heißt mit objektiv korrekten Daten „gefüttert" wurde und der Algorithmus diese richtig um Sinne des Behandlungsziels auswertet. Ob dies der Fall ist, ist für den Verwender jedoch im Regelfall nicht ersichtlich.[59] Liegen allerdings Anzeichen dafür vor, dass eine KI-Software fehlerhaft ist, so wäre eine Verwendung pflichtwidrig.[60] Die bloße Verwendung einer solchen Software stellt für sich allerdings – wie oben festgestellt – noch keine Pflichtverletzung dar.[61]

Es stellt sich jedoch die Frage, ob jedwede Software eingesetzt werden darf oder dem Anwender gewisse Grenzen gesetzt sind. So stellt KI-Software, die zur Behandlung oder Diagnostik genutzt wird, ein Medizinprodukt im Sinne des Art. 2 Nr. 1 MedProdVO (EU) 2017/ 745 dar. Eine solche Software muss daher den Anforderungen an Mangelfreiheit des Art. 5 Abs. 1 MedProdVO (EU) 2017/ 745 sowie den grundlegenden Sicherheits- und Leistungsanforderungen des Art. 5 Abs. 2 MedProdVO (EU) 2017/ 745 an ein solches Produkt genügen und ist dementsprechend zu zertifizieren, Art. 52 MedProdVO (EU) 2017/ 745.[62] Diese regulatorischen Anforderungen an KI-Software als Medizinprodukt haben auch einen Einfluss auf die Arzthaftung. So ist bei der Nutzung einer nicht zertifizierten KI-Software zur Behandlung oder Diagnose von einer Pflichtverletzung auszugehen.[63] Die Anforderung zur Nutzung eines zertifizierten Medizinprodukts stellt jedoch freilich kein Novum dar, sondern vielmehr eine klassische Voraussetzung für die Nutzung eines Medizinprodukts. Auf der anderen Seite wird man annehmen können, dass ein Nutzer von einer grundsätzlich korrekten Funktionsweise eines Programms wird ausgehen dürfen, wenn es entsprechend zertifiziert wurde.

---

[57] Siehe bspw. zu dem Programm Watson von IBM: *Jörg*, Digitalisierung in der Medizin, S. 89.
[58] *Ernst*, Rechtsfragen der Systemmedizin, S. 138ff.
[59] *Bördner*, GuP 2019, 131, 133f; *Helle*, MedR 2020, 993, 998.
[60] *Helle*, MedR 2020, 993, 998.
[61] Siehe C.I.1.
[62] *Ströbel/Grau*, ZD 2022, 599, 603ff.; *Armbrüster/Prill*, ZVersWiss 2022, 177, 182.
[63] *Helle*, MedR 2020, 993, 998; *Dettling*, PharmR 2019, 633, 641; *Armbrüster/Prill*, ZVersWiss 2022, 177, 182.

b. Der Einfluss von KI-Software auf den „ärztlichen Standard“

Aufgrund der enorm großen Menge an Daten, die intelligente Algorithmen auszuwerten in der Lage sind, hat KI-Software das Potenzial, das ärztliche Arbeiten in vielen Bereichen grundsätzlich zu verändern. Die heute schon vorhandene Menge an Daten führt dazu, dass es dem einzelnen Arzt nicht mehr möglich ist, den Stand der aktuellen Forschung genau zu überblicken.[64] Daher gibt es medizinische Leitlinien, die Handlungsempfehlungen aussprechen.[65] Leitlinien sind jedoch meist darauf ausgerichtet, allgemeine Handlungsempfehlungen für eine große Masse von Individuen zu geben.[66] Demgegenüber haben es sich Forschungsrichtungen wie die Systemmedizin zum Ziel gesetzt, durch die Nutzung u.a. von Systembiologie und Informatik Diagnose- und Therapiemethoden zu verbessern und auf das einzelne Individuum anzupassen.[67] Hierbei kommt der Nutzung von Algorithmen eine sehr große Bedeutung zu.[68] Ärzte sollen neben ihrer klinisch-individuellen Erfahrung auch befähigt werden, auf eine enorme Fülle von externen Erfahrungswerten zurückgreifen, die ihm durch den Algorithmus aufbereitet werden.

Es stellt sich dabei jedoch die Frage, in welchem Verhältnis von einem Algorithmus gegebenen Behandlungsempfehlungen und Diagnosen zum „ärztlichen Standard“ als relevantem Kriterium für die ärztliche Haftung stehen. Geht man von einer signifikanten Erhöhung der Evidenz der Behandlungsvorschläge und Diagnostik durch eine KI-Software aus, weil diese aus einer großen Anzahl von Daten Informationen ziehen und daraus lernen konnte, stellt sich die Frage, ob eine solche Software nicht selbst zum „ärztlichen Standard“ werden könnte.

Laut dem BGH soll der „Standard“ folgendes darstellen:

„Er repräsentiert den jeweiligen Stand der naturwissenschaftlichen Erkenntnisse und der ärztlichen Erfahrung, der zur Erreichung des ärztlichen Behandlungsziels erforderlich ist und sich in der Erprobung bewährt hat.“[69]

Ein Algorithmus, der auf Daten zugreifen kann, die den aktuellen Stand der naturwissenschaftlichen Erkenntnisse zu einer bestimmten Frage widerspiegelt und auch über eine Vielzahl von Patientendaten verfügt, die er miteinander ins Verhältnis setzen kann, verfügt somit über zwei der vom BGH genannten Voraussetzungen für den medizinischen „Standard“. Der Nutzung von Algorithmen fehlt zurzeit jedoch die dritte Voraussetzung, nämlich die Bewährung durch

---

[64] *Woopen*, in: Jansen/Katzenmeier/Woopen, Medizin und Standard, S. 119, 124.
[65] Siehe B.I.
[66] *Hart*, MedR 1998, 8, 9.
[67] *Ernst*, Rechtsfragen der Systemmedizin, S. 5ff; *Erdmann/Fischer/Fleßa/Langanke*, Deutsches Ärzteblatt 2015, A 1330ff.
[68] *Katzenmeier*, MedR 2021, 859.
[69] BGH NJW 2016, 713, 714; BGHZ 102, 17 [24 f.] = NJW 1988, 763; NJW-RR 2014, 1053 = VersR 2014, 879 Rn. 11.

Erprobung. Allerdings könnte man dies auch anzweifeln, da ein Algorithmus keine Art einer Behandlung ist, sondern vielmehr eine bestimmte Art der Wissensaufbereitung. Vor dem Hintergrund des „Intelligenzrisikos“[70] wird man jedoch davon ausgehen müssen, dass eine KI-Software erst selbst zum „Standard“ werden kann, wenn sie erprobt wurde.[71]

Dass KI-Software selbst jedoch das Potenzial hat, durch ihre bessere Differenzierung von Patientengruppen, der Fähigkeit zu lernen und der Analyse einer großen Menge von Daten hat, den „ärztlichen Standard“ zu prägen, ist anzunehmen.[72] Allerdings unterscheidet sich die Vorgehensweise einer KI-Software in vielen Fällen von dem Konzept des „ärztlichen Standards“ und ärztlicher Leitlinien. Wo ein „Standard“ für eine große Anzahl gleich gelagerter Fälle gelten soll, sind die Behandlungsempfehlungen eines Algorithmus auf wesentlich kleinere, differenziertere Patientengruppen zugeschnitten.[73] Es findet eine Ausdifferenzierung statt, in der wesentlich mehr Bedingungen betrachtet werden, als bei einer klassischen Standardbildung.[74] Wenn man nun davon ausgeht, dass KI-Software das Potenzial hat, durch differenzierte Diagnosen selbst zum „Standard“ zu werden, jedoch die Standardbildung eines Algorithmus nicht mit der Bildung eines klassischen „Standards“ vergleichbar ist, so stellt sich die Frage, ob nicht das Vorgehen bei der Standardbildung generell vor dem Hintergrund der neueren technischen Entwicklungen angepasst werden muss. Dabei wird vorgeschlagen, den Standardbegriff nicht mehr krankheitsbezogen zu formulieren, sondern im konkreten Bezug auf einzelne Patientengruppen.[75] Dies hätte zur Folge, dass es weit mehr und individuellere „Standards“ als heute geben würde und der Standardbegriff somit eine Wandlung erfahren würde.[76] Voraussetzung hierfür wird jedoch sein, dass sich Systeme, die auf selbstlernenden Algorithmen basieren, auch in der Praxis flächendeckend durchsetzen.

Wenn – wie oben festgestellt – die grundsätzliche Möglichkeit besteht, dass die Nutzung von KI-Software selbst zum „Standard“ werden kann, so werden sich in Zukunft unter dieser Prämisse zum Teil andere Haftungsfragen stellen als heute. Von wesentlicher Relevanz ist dabei die Frage, welche haftungsrechtlichen Folgen es haben kann, wenn eine KI-Software nicht verwendet wird, obwohl ihre Nutzung zum „Standard“ geworden ist. In der Literatur wird davon ausgegangen, dass dies grundsätzlich eine Pflichtverletzung darstellen kann.[77] Allerdings ist

---

[70] Siehe A.III.3.
[71] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 13.
[72] *Helle*, MedR 2020, 993, 998; *Ernst*, Rechtsfragen der Systemmedizin, S. 18; *Kienzle*, in: Jansen/Katzenmeier/Woopen, Medizin und Standard, S. 37, 42f.; *Dettling*, PharmR 2019, 633, 642; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 197.
[73] *Ernst*, Rechtsfragen der Systemmedizin, S. 139; *Keil*, Rechtsfragen, S. 139.
[74] *Woopen*, in: Jansen/Katzenmeier/Woopen, Medizin und Standard, S. 119, 125.
[75] *Hart*, MedR 2016, 669, 674; *Ernst*, Rechtsfragen der Systemmedizin, S. 150.
[76] Ebd.
[77] *Hart*, MedR 2016, 669, 672; *Ernst*, Rechtsfragen der Systemmedizin, S. 150; *Katzenmeier*, MedR 2019, 259, 268; *Taupitz*, AcP 211 (2011), 352, 386.

hier eine differenzierte Betrachtung geboten. Der BGH hat zu der Einhaltung von „Standards“ folgendes geäußert:

„Der Arzt schuldet seinem Patienten neben einer sorgfältigen Diagnose die Anwendung einer Therapie, die dem jeweiligen Stand der Medizin entspricht. ***Indessen bedeutet das nicht, daß jeweils das neueste Therapiekonzept verfolgt werden muß, wozu dann auch eine stets auf den neuesten Stand gebrachte apparative Ausstattung gehören müßte.*** *Der Zeitpunkt, von dem ab eine bestimmte Behandlungsmaßnahme veraltet und überholt ist, so daß ihre Anwendung nicht mehr dem einzuhaltenden Qualitätsstandard genügt und damit zu einem Behandlungsfehler wird, ist jedenfalls dann gekommen, wenn neue Methoden risikoärmer sind und/oder bessere Heilungschancen versprechen, in der medizinischen Wissenschaft im wesentlich unumstritten sind und deshalb nur ihre Anwendung von einem sorgfältigen und auf Weiterbildung bedachten Arzt verantwortet werden kann (Deutsch, ArzthaftungsR und ArzneimittelR, S. 35).* ***Da aber schon aus Kostengründen, anfangs möglicherweise auch wegen eines noch unzureichenden Angebotes auf dem Markt, nicht sofort jede technische Neuerung, die den Behandlungsstandard verbessern kann, von den Kliniken angeschafft werden kann, muß es für eine gewisse Übergangszeit ferner gestattet sein, nach älteren, bis dahin bewährten Methoden zu behandeln, sofern das nicht schon wegen der Möglichkeit, den Patienten an eine besser ausgestattete Klinik zu überweisen, unverantwortlich erscheint.“***[78]

Es ist somit nicht zwangsläufig immer die neueste und einen höheren Erfolg versprechende Behandlung geschuldet, sondern vielmehr eine solche, die noch zum „Standard“ gehört.[79] Für die Frage, zu welchem Zeitpunkt man davon ausgehen können wird, dass die Nichtnutzung einer KI-Software eine Pflichtverletzung darstellt, wird somit entscheidend sein, wie sehr sie sich in der klinischen Praxis durchsetzen und sich dabei als überlegen gegenüber anderen Methoden erweisen wird. Zum aktuellen Zeitpunkt, zu dem die Nutzung von KI-Software sich noch nicht flächendeckend durchgesetzt hat und sich eher in einem Erprobungsstadium befindet, ist nicht davon auszugehen, dass eine Nichtnutzung solcher Software eine Pflichtverletzung darstellt.[80]

c. Fazit

Es kann somit festgehalten werden, dass die Nutzung von KI-Software das grundsätzliche Potenzial hat, selbst zum „Standard“ zu werden. Allerdings wird dies erst der Fall sein, wenn sich diese Programme in der Praxis bewährt und durchgesetzt haben. Weiterhin sind die

[78] BGH NJW 1988, 764, 764ff.
[79] BGH NJW 1992, 754.
[80] so auch: *Spindler*, in: Katzenmeier (Hrsg.), FS Hart S. 581, 587.

grundsätzlichen regulatorischen Vorgaben zu beachten, sodass nur ein Medizinprodukt auch zum „Standard“ werden kann, dass zertifiziert wurde.

Sollte die Nutzung von KI-Software in bestimmten Bereichen zum „Standard“ werden, so würde sich in diesen aller Voraussicht nach auch die Art des „Standards“, weg von der Empfehlung von Behandlungsmethoden für Krankheitsbilder, hin zu Handlungsempfehlungen bestimmter (kleiner) Patientengruppen wandeln. Der „Standard“ würde damit wesentlich individualisierter auf den einzelnen Patienten zugeschnitten sein. Dies wird auch noch nicht absehbare Folgen auf das Arzthaftungsrecht und die Rolle des „Standards“ bei Feststellung einer Pflichtverletzung haben.

Sollte die Nutzung von KI-Software sich in bestimmten Bereichen langfristig durchgesetzt haben und zum „Standard“ geworden sein, so kann auch eine Nichtnutzung grundsätzlich eine Pflichtverletzung darstellen. Allerdings würde dies nicht nur voraussetzen, dass KI-Software zum „Standard“ würde, sondern auch, dass ihre Nutzung zur Behandlung anderen Methoden so überlegen wäre, dass diese nicht mehr guten Gewissens angewandt werden dürften. Grundsätzlich ist aber nach der Rechtsprechung des BGH vorstellbar, dass, selbst wenn die Nutzung einer KI-Software in einem Bereich der „Standard“ geworden ist, andere (schlechtere) Methoden für einen Übergangszeitraum noch angewendet werden dürfen.

3. Anforderungen an die ärztliche Sorgfalt bei der Nutzung von KI-Software

a. Anforderungen an den Arzt

DieAnforderungen an den Arzt bei der Verwendung von einer KI-Software für Diagnostik und Behandlung unterscheiden sich nicht grundsätzlich von denen beim Einsatz anderer technischer Geräte. Hierzu hat der BGH schon im Jahre 1977 geschrieben:

„Zwar bringt es die zunehmende Technisierung der modernen Medizin mit sich, daß der Arzt nicht mehr alle technischen Einzelheiten der ihm verfügbaren Geräte zu erfassen und gegenwärtig zu haben vermag (vgl. Senat, NJW 1975, 2245 = VersR 1975, 952 [953]). Das befreit ihn aber nicht von der Pflicht, sich mit der Funktionsweise insbesondere von Geräten, deren Einsatz für den Patienten vitale Bedeutung hat, wenigstens insoweit vertraut zu machen, wie dies einem naturwissenschaftlich und technisch aufgeschlossenen Menschen (diese Fähigkeiten müssen vor allem bei einem Anästhesisten vorausgesetzt werden) möglich und zumutbar ist.“[81]

[81] BGH Urt. v. 11.10.1977 – VI ZR 110/75, NJW 1978, 584, 585.

Von dem Verwender ist somit zu erwarten, dass er sich mit der grundsätzlichen Funktionsweise der genutzten KI-Software vertraut macht. Dies bedeutet jedoch nicht, dass vertiefte Informatikkenntnisse erworben werden müssten. Vielmehr reicht es aus, wenn der Nutzer sich mit der Funktionsweise des selbstlernenden Algorithmus und der Datenbasis auseinandersetzt. Gerade bei der Datenbasis ist dabei von einem kundigen Anwender zu fordern, dass er hier besonders umsichtig ist. Auch ein selbstlernender Algorithmus kann nur mit den Daten arbeiten, die ihm zur Verfügung gestellt werden. Wenn eine neue, sehr effiziente Therapiemethode in den vom Algorithmus genutzten Daten nicht auftaucht, so entsteht bei der Auswahlentscheidung einer Therapiemethode für die KI an dieser Stelle ein blinder Fleck. Hier ist es Aufgabe des Arztes darauf zu achten, dass ein Programm auf dem aktuellen Stand bleibt, und seine Ergebnisse kritisch vor dem Hintergrund aktueller medizinischer Entwicklungen auf Plausibilität zu prüfen. Gleichzeitig trifft den Verwender die Verantwortung, das Programm regelmäßig zu warten,[82] was im Falle einer Software bedeutet, die angebotenen Updates, Patches und Bugfixes umgehend zu installieren.

Teilweise wird angenommen, dass sich die Sorgfaltspflichten des Arztes aufgrund des „Intelligenzrisikos“[83] verdichten würden.[84] Fraglich ist allerdings, wie die statuierten besonderen Sorgfaltspflichten des Verwenders von KI-Software aufgrund des „Intelligenzrisiko“ aussehen sollten. Gerade die fehlende (vollständige) Nachvollziehbarkeit der Entscheidungen lassen besondere Überprüfungspflichten ins Leere laufen. Es bleibt vielmehr bei der grundsätzlichen Verpflichtung, sich mit der genutzten Software vertraut zu machen und diese auf dem aktuellen Stand zu halten. Ein Unterschied zu der Nutzung anderer technischer Geräte besteht trotz des „Intelligenzrisikos“[85] de lege lata jedoch nicht.

b. Zurechnung von Maschinenfehlern über § 278 BGB

Nach dem bisher gesagten haftet der Verwender einer KI-Software nicht für deren Fehler, wenn er diese nicht vorhersehen konnte, was bei selbstlernenden Algorithmen nur äußerst selten der Fall sein dürfte. Es wird daher eine Haftungslücke zulasten der Patienten

[82] *Katzenmeier*, MedR 2021, 859, 860f.; MüKoBGB-*Wagner*, Vor § 630a Rn 60; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 40.
[83] Siehe A.III.3.
[84] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 41.
[85] Siehe A.III.3.

befürchtet.[86] Nach teilweise vertretener Ansicht soll diese mittels einer analogen Anwendung des § 278 BGB geschlossen werden.[87]

Das Verschulden einer anderen Person (dem Erfüllungsgehilfen) ist nach dieser Norm dem Schuldner zurechenbar, wenn er sich dieser zur Erfüllung seiner Verbindlichkeit bedient. Bei einer KI-Software handelt es sich zwar um keine Person[88], allerdings wird eine analoge Anwendung des § 278 BGB ins Spiel gebracht, weil ansonsten das Ausgleichsinteresse von Geschädigten nicht ausreichend berücksichtigt würde und es hierdurch bei der Nutzung von KI-Software bei Behandlung und Diagnostik im Schadensfall zu Haftungslücken kommen könnte.[89] Selbst, wenn man davon ausgeht, dass § 278 BGB analog auf künstliche Intelligenzen angewendet werden kann, fehlt es dieser jedoch mangels Verantwortungsbewusstsein an einer subjektiven Verschuldenskomponente.[90] Hiergegen wird zwar eingewandt, dass der Fahrlässigkeitsbegriff im Zivilrecht stark objektiviert sei.[91] Doch auch wenn bei einfacher Fahrlässigkeit der objektive Sorgfaltsmaßstab regelmäßig entscheidend ist,[92] bedeutet dies nicht, dass subjektive Komponenten gar keine Rolle mehr spielen würden. Dies zeigt sich bspw. bei der Berücksichtigung von medizinischen Spezialkenntnissen im Rahmen der Fahrlässigkeitsprüfung.[93] Ein zurechenbares Verschulden einer KI-Software kommt vor diesem Hintergrund nicht in Betracht. Hierfür spricht zudem auch, dass zur Bestimmung des objektiven Sorgfaltsmaßstabs teilweise Regelwerke herangezogen werden, deren Adressaten KI-Systeme mangels einer Rechtssubjektivität nicht sein können.[94] Entstehende Haftungslücken sind daher durch den Gesetzgeber zu schließen. Dies wurde von diesem auch erkannt und mit einer Umsetzung begonnen.[95]

---

[86] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 44; *Katzenmeier*, MedR 2021, 859, 861; *Armbrüster/Prill*, ZVers Wiss 2022, 177, 182; *Teubner*, AcP 218 (2018), 155, 188ff; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 197f.
[87] *Spindler*, in: Katzenmeier (Hrsg.), FS Hart S. 581, 585; *Helle*, MedR 2020, 993, 998; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 44; *Katzenmeier*, MedR 2021, 859, 861; *Armbrüster/Prill*, ZVers Wiss 2022, 177, 182; *Teubner*, AcP 218 (2018), 155, 188ff.
[88] Vergleiche hierzu das teilweise diskutierte Konzept einer E-Person unter C.V.1.b.
[89] MüKoBGB-*Wagner*, Vor § 630a Rn 61; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 44; *Spindler*, in: Katzenmeier (Hrsg.), FS Hart S. 581, 585.
[90] *Katzenmeier*, MedR 2021, 859, 861; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 198.
[91] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 Rn 44; *Spindler*, in: Katzenmeier (Hrsg.), FS Hart S. 581, 585; *Teubner*, AcP 218, 2018, 155, 188.
[92] MüKoBGB-*Grundmann*, § 276 Rn 55.
[93] BGH NJW 1987, 1479.
[94] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 200.
[95] Siehe C.V.2.

4. Fazit

Im Hinblick auf die Haftung des Arztes aus Behandlungsvertrag lässt sich somit festhalten, dass allein die Nutzung einer KI-Software noch keine Pflichtverletzung darstellt.[96] Genauso wenig stellt die Nichtnutzung von KI zurzeit eine Pflichtverletzung dar.[97] Dies wird zumindest auch so lange so bleiben, wie die Nutzung von KI-Software im Rahmen von Behandlung und Diagnostik noch eine *Neulandmethode* darstellt.[98] Allerdings hat eine solche Software das Potenzial, in bestimmten Bereichen flächendeckende Akzeptanz zu finden, sodass ihre Anwendung selbst zum medizinischen „Standard“ wird.[99] Dies kann indes nur für KI-Software gelten, die nach den üblichen regulatorischen Verfahren zertifiziert wurde.[100] Wird KI-Software in bestimmten Bereichen zum „Standard“, so wird dies aller Voraussicht nach auch Folgen für die Standardbestimmung weg von einem krankheitsbezogenen, hin zu einem „Standard“ haben, der auf kleinere Patientengruppen bezogen ist.[101]

Bei der Nutzung von KI-Software bestehen für den Anwender Sorgfaltspflichten. Diese sind jedoch nicht grundsätzlich verschieden zu denen bei der Nutzung anderer medizinischer Gerätschaften.[102] Die teilweise vertretene Zurechnung von Fehlverhalten von KI-Software über § 278 BGB ist abzulehnen, da es KI-Software jedenfalls an einer subjektiven Verschuldenskomponente fehlt.[103] Diese subjektive Komponente kann es allenfalls bei den Entwicklern der KI geben, die aber keine Erfüllungsgehilfen des behandelnden Arztes sind.

II. Besonderheiten bei der deliktischen Haftung bei Nutzung vom KI-Software

1. Haftung nach § 823 BGB

Hinsichtlich der deliktischen Arzthaftung nach § 823 BGB ergeben sich keine wesentlichen Besonderheiten gegenüber der Vertragshaftung, da beide im Arzthaftungsrecht weitgehend

[96] Siehe C.I.1.
[97] Siehe C.I.2.b.
[98] Siehe C.I.2.a.
[99] Siehe C.I.2.b.
[100] Siehe C.I.2.a.
[101] Siehe C.I.2.b.
[102] Siehe C.I.3.a.
[103] Siehe C.I.3.b.

gleichlaufend sind.[104] Der Verwender von KI-Software haftet somit grundsätzlich für mangelnde Wartung, Instandhaltung und Fehlnutzung von KI.[105]

2. Haftung nach § 831 BGB für KI-Software als digitalem Verrichtungsgehilfen

Diskutiert wird eine Haftung für KI-Software nach § 831 BGB analog.[106] Hier besteht ein Unterschied zu der bereits erörterten Zurechnung der Pflichtverletzung eines Erfüllungsgehilfen über § 278 BGB in der Form, dass bei der Verpflichtung zum Ersatz des Schadens durch einen Verrichtungsgehilfen nach § 831 Abs. 1 BGB nicht auf ein Verschulden ankommt.[107] Das Problem des Bestehenmüssens einer subjektiven Verschuldenskomponente wie bei der Haftung des Erfüllungsgehilfen[108] stellt sich demnach nicht. Das Fehlen einer solchen subjektiven Komponente führt dazu, dass auch KI-Software, die kein Rechtssubjekt darstellt, als „digitaler Verrichtungsgehilfe“ angesehen werden könne.[109] KI fehle es zwar an „responsibility“ für ihr Handeln, nicht aber an „accountability“, weswegen sie rechtswidrig handeln könne.[110] Hiergegen wird jedoch eingewandt, dass ein nicht rechtsfähiges Objekt wie eine KI-Software, die nicht Träger von Rechten und Pflichten sein könne, auch nicht widerrechtlich handeln könne.[111] Zudem wird angemerkt, dass § 831 BGB rechtsdogmatisch zwei Schuldner, nämlich den Geschäftsherren und den Verrichtungsgehilfen vorsehe. Dies sei bei einem „digitalen Verrichtungsgehilfen“ jedoch nicht gegeben.[112]

So viel Uneinigkeit im Hinblick auf das Konstrukt des „digitalen Verrichtungsgehilfen“ herrscht, herrscht umso mehr Einigkeit im Hinblick auf die Geeignetheit der analogen Anwendung von § 831 BGB, um Haftungslücken beim Einsatz von KI Software zu schließen. Nach § 831 Abs. S. 2 BGB besteht für den Geschäftsherren eine Exkulpationsmöglichkeit im Hinblick auf Pflichtverletzungen seines Verrichtungsgehilfen durch den Nachweis der sorgfältigen Auswahl und Leitung des Verrichtungsgehilfen. Es reicht mithin der Nachweis des Verwenders aus, dass er seine Sorgfaltspflichten erfüllt und das System den Schaden eigenständig verursacht hat oder aber, dass der Schaden auch bei nicht sorgfaltspflichtwidrigen Verhalten eingetreten wäre.[113]

---

[104] *Katzenmeier*, MedR 2021, 859, 861.
[105] MüKoBGB-*Wagner*, Vor § 630a Rn 63.
[106] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 202; MüKoBGB-*Wagner*, Vor § 630a Rn 62; *Katzenmeier*, MedR 2021, 859, 861f.; *Brand*, MedR 2019, 943, 949.
[107] BeckOKBGB-*Förster* § 831 Rn 35; MüKoBGB-*Wagner*, § 831 Rn 34.
[108] Siehe C.I.3.b.
[109] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 202; MüKoBGB-*Wagner*, Vor § 630a Rn 62; *Katzenmeier*, MedR 2021, 859, 861f.; *Brand*, MedR 2019, 943, 949.
[110] *Katzenmeier*, MedR 2021, 859, 862.
[111] *Brand*, MedR 2019, 943, 949; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 204.
[112] *Brand*, MedR 2019, 943, 949.
[113] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 204.

Es wäre daher auch bei einer analogen Anwendung nicht zu erwarten, dass hierdurch Haftungslücken voll umfassend geschlossen werden könnten.[114]

Allerdings ist aus hiesiger Lesart bereits die analoge Anwendung des § 831 auf „digitale Verrichtungsgehilfen“ abzulehnen, da sie aus oben genannten Gründen rechtsdogmatisch höchst fragwürdig ist.

3. Analoge Anwendung der Tierhalterhaftung § 833 BGB

Eine weitere Überlegung geht in die Richtung, die Tierhalterhaftung nach § 833 BGB auf die Nutzung von KI-Software zu übertragen.[115] Hiernach könnte – je nachdem, ob man bei einer KI-Software die Analogie zu einem Luxus- oder Nutztier bejahen würde – eine Gefährdungshaftung bzw. eine Haftung für vermutetes Verschulden des Halters bzw. Verwenders treten.[116] Durch die analoge Anwendung könnten die Haftungslücken geschlossen werden, die aufgrund der regelmäßig nicht möglichen Verschuldenszurechnung bei der Nutzung von KI-Software auftreten.[117] Die Befürworter dieser Ansicht betonen, dass die Regelungen zur Tierhalterhaftung deswegen bestünden, da von einem Tier unkontrollierbare Instinkthandlungen ausgehen würden, denen eine besondere Gefährlichkeit innewohnen würde. Gleiches würde auch für KI-Software gelten, denn auch deren Verhalten sei unvorhersehbar.[118] Als weiterer (praktischer) Aspekt wird angeführt, dass eine solche Haftung wohl von bestehenden Haftpflichtversicherungspolicen gedeckt sei.[119]

Gegen eine analoge Anwendung spricht jedoch, dass eine Gefährdungshaftung im deutschen Recht die Ausnahme ist und daher nur begrenzt analogiefähig ist.[120] Zudem besteht die teilweise angenommene Vergleichbarkeit zwischen Tieren und KI-Software nur in Teilen. Zwar ist das Verhalten von KI-Software für den Menschen nicht immer nachvollziehbar, allerdings handelt es sich, im Gegensatz zu tierischen Verhalten, gerade nicht um Instinkthandlungen. Vielmehr entscheidet die Software nach rationalen Kriterien.[121] Dogmatisch gesehen ergibt es zudem wenig Sinn, wenn, der Systematik von § 833 BGB folgend, die private Nutzung von KI-Software nach § 833 S. 1 BGB und die geschäftliche Nutzung nach § 833 S. 2 BGB

---

[115] *Riehm/Meier*, in Fischer/Hoppen/Wimmers, DRGI Jahrbuch 2018, Rn 25; *Borges*, NJW 2018, 977, 981; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 45.
[116] BeckOKBGB-*Spindler*, § 833 Rn 1; Staudinger-*Eberl-Borges*, § 833 Rn 5ff.
[117] C.I.3.b.
[118] *Zech*, ZfPW 2019, 198, 214; *Brand*, MedR 2019, 943, 949; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 208.
[119] *Brand*, MedR 2019, 943, 949.
[120] *Katzenmeier*, MedR 2021, 859, 862; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 208.
[121] Ebd.

ausgestaltet wäre. Dies würde dazu führen, dass der Einsatz von KI-Software im medizinischen Bereich, gegenüber dem rein privaten Gebrauch, für die dann die Gefährdungshaftung greifen würde, bessergestellt würde.[122]

Vor diesem Hintergrund kommt auch eine analoge Anwendung der Tierhalterhaftung nach § 833 BGB nicht infrage.[123]

4. Analogien zu § 832 und § 836 BGB

Darüber hinaus wird vereinzelt eine analoge Anwendung von § 832 BGB diskutiert, welcher die Haftung des Aufsichtspflichtigen normiert. Allerdings wird dies zu Recht mit Verweis auf die mangelnde Analogiefähigkeit der Norm sowie aus dogmatischen Erwägungen abgelehnt.[124]

Gleiches gilt für eine analoge Anwendung des § 836 BGB, der die Haftung des Grundstücksbesitzers wegen des besonderen Gefahrenpotenzials von Gebäuden normiert. Eine solche scheitert an der für eine analoge Anwendung notwendigen vergleichbaren Interessenlage.[125]

5. Fazit

Hinsichtlich der deliktischen Haftung aus § 823 ergeben sich auch beim Einsatz von KI-Software keine Besonderheiten.[126] Die in der Literatur teilweise angenommenen Möglichkeiten zur analogen Anwendung von §§ 831, 833, 832 und 836 BGB sind abzulehnen.[127] Vielmehr ist es am Gesetzgeber, rechtliche Lösungen für etwaige Haftungslücken zu schaffen.

III. Der Einfluss der Nutzung von KI-Software auf die Aufklärungspflicht nach § 630e BGB

Nach § 630e BGB besteht eine Aufklärungspflicht des Behandelnden gegenüber dem Patienten. Diese bezieht sich nach § 630e Abs. 1 S. 2 BGB auf Art, Umfang, Durchführung, zu erwartende Folgen, Risiken der Maßnahme sowie Notwendigkeit, Dringlichkeit, Eignung und Erfolgsaussichten im Hinblick auf die Diagnose oder Therapie. Die Verletzung der Aufklärungspflicht stellt eine Pflichtverletzung des Behandlungsvertrags dar und es entfällt hierdurch ggf.

---

[122] *Borges*, NJW 2018, 977, 981.
[123] Andere Ansicht mit detaillierter Begründung siehe: *Brand*, MedR 2019, 943, 949.
[124] *Brand*, MedR 2019, 943, 949; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 205.
[125] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 209; *Borges*, NJW 2018, 977, 981.
[126] Siehe C.II.1.
[127] Siehe C.II.2-4.

die Rechtfertigungswirkung der Einwilligung in die Verletzung des Körpers des Patienten durch den Behandler.[128]

Bei der Anwendung einer *Neulandmethode* ist der Patient grundsätzlich darüber aufzuklären, dass es sich um eine solche handelt.[129] Zudem besteht die Verpflichtung, über die Vor- und Nachteile der *Neulandmethode* und bekannte Risiken aufzuklären sowie darüber zu informieren, dass nicht auszuschließen ist, dass noch unbekannte Risiken bestehen.[130] Der Patient muss nach dem BGH daher in die Lage versetzt werden, selbst aufgrund einer vollumfänglichen Information über Risiken und Vorteile entscheiden zu können, nach welcher Methode er behandelt wird.[131] Solange KI-Software eine *Neulandmethode* darstellt, ist somit in diesem Sinne über ihre Verwendung aufzuklären.[132] Dabei hat sich die Aufklärung vor allen Dingen auch auf die fehlende Beherrschbarkeit der KI, also das „Intelligenzrisiko“[133], zu beziehen.[134]

Sollte die Verwendung von KI-Software zum „ärztlichen Standard“ werden,[135] so gelten die allgemeinen Regeln über den Umfang der Aufklärung. Danach muss der Patient „im Großen und Ganzen“ über die Schwere und Folgen der Behandlung aufgeklärt werden.[136] Fraglich ist jedoch, ob in einem solchen Fall gleichwohl noch stets über das „Intelligenzrisiko“[137] aufzuklären ist.[138] Jedenfalls bei einer Behandlung, die ein hohes Gesundheitsrisiko birgt, wird dies wohl zu bejahen sein.[139] Da der Einsatz von KI-Software sich zurzeit noch nicht breit durchgesetzt hat, sollte über ihren Einsatz grundsätzlich aufgeklärt werden, um Haftungsrisiken aus dem Weg zu gehen.[140]

Die Aufklärung eines Patienten hat nach § 630e Abs. 2 Nr. 1 1. HS BGB mündlich durch den Behandelnden oder durch eine Person zu erfolgen, die über die zur Durchführung der Maßnahme notwendige Ausbildung verfügt. Eine Aufklärung des Patienten durch KI-Software ist damit de lege lata unzulässig.[141]

---

[128] Staudinger-*Gutmann*, § 630e, Rn 186; MüKoBGB-*Wagner* § 630e, Rn 89ff.
[129] Staudinger-*Gutmann*, § 630e, Rn 34; MüKoBGB-*Wagner* § 630e, Rn 37ff.
[130] Staudinger-*Gutmann*, § 630e, Rn 34 m.w.N.
[131] BGHZ 169, 103 = NJW 2006, 2477 Rn 14.
[132] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 16f.; *Spindler*, FS Hart S. 581, 592; *Katzenmeier*, MedR 2021 859, 861; *Ernst*, S. 156.
[133] Siehe A.III.3.
[134] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 17; *Katzenmeier*, MedR 2021 859, 861;
[135] Siehe C.I.2.b.
[136] Staudinger-*Gutmann*, § 630e, Rn 16; MüKoBGB-Wagner § 630e, Rn 17; Chibanguza/Kuß/Steege -*Eichelberger*, KI, § 4 I Rn 18.
[137] Siehe A.III.3.
[138] *Katzenmeier*, MedR 2021, 859, 861; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 18ff.
[139] Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 19; *Dettling*, PharmR 2019, 633, 641.
[140] Siehe auch: Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 22
[141] *Spindler*, FS Hart S. 581, 590; Katzenmeier, MedR 2021, 859, 861; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 24f.

Neben der Verpflichtung zur Aufklärung des Patienten über die Behandlungsmethode, stellt sich beim Einsatz von KI-Software auch die Frage, ob es nicht gegebenenfalls eine Pflicht des Arztes gibt, über bestimmte, mittels der KI-Software erworbene Erkenntnisse, nicht zu informieren, mithin also, ob der Patient ein subjektives Recht auf Nichtwissen hat.[142] Durch die Fähigkeit von KI-Software, Datensätze umfassend auszuwerten und Verknüpfungen zu ziehen, ergeben sich gerade im Bereich der Diagnostik vielfältige neue Ansatzpunkte, wobei das Wissen über bestimmte gesundheitliche Dispositionen oder die Zugehörigkeit zu einer Risikogruppe für den Patienten zu einer Belastung werden kann.[143] Allerdings wurde durch den BGH in der Vergangenheit eine Persönlichkeitsrechtsverletzung in Form des „Rechts auf Nichtwissen" in einem Fall des Hinweises auf die Trägerschaft einer Erbkrankheit angezweifelt, da *„eine freie Entscheidung, bestimmte Informationen nicht erhalten zu wollen, voraussetzt, dass der Betroffenen weiß, dass es Informationen gibt, die er zur Kenntnis nehmen könnte."* Es steht jedoch zu erwarten, dass es hier im Hinblick auf die neuen Möglichkeiten von Big Data zu einer Vertiefung der Diskussionen hierzu kommt.

IV. Der Einfluss der Nutzung von KI-Software auf die Beweislast nach § 630h BGB

Das deutsche Beweisrecht fußt auf dem Grundsatz, dass derjenige, der aus einer für ihn günstigen Norm Rechte herleitet, deren Voraussetzungen auch darzulegen und zu beweisen hat.[144] Die Haftung aus Behandlungsvertrag ist im Hinblick auf die Beweislast jedoch Sonderregelungen unterworfen. Diese wurden durch die Rechtsprechung entwickelt und durch den Gesetzgeber in § 630h BGB verankert.[145] Nach § 630h Abs. 1 BGB wird ein Fehler des Behandelnden vermutet, *wenn sich ein allgemeines Behandlungsrisiko verwirklicht hat, das für den Behandelnden voll beherrschbar war und das zur Verletzung des Lebens, des Körpers oder der Gesundheit des Patienten geführt hat.*

Bei Schäden durch den Einsatz von KI-Software stellt sich die Frage, ob die Beweislastumkehr des § 630h Abs. 1 BGB greifen kann, da sich schließlich ein Behandlungsrisiko verwirklicht haben muss, dass für den Behandelnden *voll beherrschbar* war. Bei der Nutzung technischer Geräte wird grundsätzlich davon ausgegangen, dass deren Funktionsfähigkeit und ordnungsgemäße Bedienung zu den voll beherrschbaren Risiken gehört.[146] Allerdings stellt sich die

---

[142] Siehe: *Hahn*, MedR 2019, 197ff.
[143] *Ernst*, Rechtsfragen der Systemmedizin S. 157; *Hahn*, MedR 2019, 197.
[144] Staudinger-*Gutmann* § 630h Rn 3.
[145] Baumgärtel/Laumen-*Repgen*, § 630h Rn 1; Staudinger-Gutmann § 630h Rn 13.
[146] Baumgärtel/Laumen-*Repgen*, § 630h Rn 57ff.; *Spindler*, FS Hart S. 581, 593; Chibanguza/Kuß/Steege-*Eichelberger*, § 4 I Rn 49.

Frage, ob dies auch für den Einsatz von KI-Software mit selbstlernenden Algorithmen übertragbar ist. Aufgrund des „Intelligenzrisikos“[147] kann der Verwender von KI-Software diese nämlich gerade nicht voll beherrschen.[148] Nach dem insoweit klaren Wortlaut der Norm greift § 630h BGB daher bei KI-Software nicht, wenn der Verwender nicht mehr in der Lage ist, das Vorgehen der Software vollumfänglich verstehen zu können.[149] Teilweise wird eine teleologische Reduktion des § 630h Abs. 1 im Hinblick auf die volle Beherrschbarkeit des Risikos vorgeschlagen, da Sinn und Zweck der Norm eine gerechte Risikozuordnung in die jeweilige Herrschaftsspähre sei.[150] Dies ist jedoch im Hinblick auf den eindeutigen Wortlaut abzulehnen. Vielmehr dürfte auch hier der Gesetzgeber gefordert sein, die Norm im Hinblick auf den Einsatz von KI-Software im medizinischen Bereich einer kritischen Prüfung zu unterziehen.[151]

V. Perspektiven auf die zukünftige Rechtsentwicklung

Sowohl in der Literatur, als auch in der Gesetzgebung gibt es Ansätze, regulatorisch auf die zunehmende Nutzung von KI-Software zu reagieren. Im Mittelpunkt steht hier häufig die Frage, ob für Schäden, die auf der Anwendung von KI beruhen, eine Haftung eher auf Hersteller- oder Betreiberseite zu sehen ist.[152]

1. Vorschläge aus der Literatur

In der Literatur werden unterschiedliche Vorschläge im Hinblick auf das Schließen etwaiger Haftungslücken diskutiert.[153]

a. E-Person

Der diesem Konzept zugrundeliegende Gedanke ist, autonom agierende Systeme als E-Personen und somit als eigene Rechtspersönlichkeiten zu behandeln.[154] Dies wird jedoch aus unterschiedlichen Gründen überwiegend abgelehnt. Neben der Kritik an einem derart schweren Eingriff in die Rechtsdogmatik[155], wird ein solches Konzept auch aus ethischen Gesichtspunkten kritisch betrachtet.[156] Auf der Rechtsfolgenseite stellt sich zudem die Frage, was eine

[147] Siehe A.III.3.
[148] *Spindler*, FS Hart S. 581, 593; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 49.
[149] Ebd.
[150] *Brand*, MedR 2019, 943, 950.
[151] *Spindler*, FS Hart S. 581, 594.
[152] *Zech/Hünefeld*, MedR 2023, 1, 6.
[153] MüKoBGB-*Wagner*, Vor § 630a Rn 62; *Mühlböck/Taupitz*, AcP 221 (2021), 179, 213ff.; *Brand*, MedR 2019, 943, 947; Chibanguza/Kuß/Steege-*Eichelberger*, KI, § 4 I Rn 45.
[154] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 213f.; *Brand*, MedR 2019, 943, 947f.
[155] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 215; *Brand*, MedR 2019, 943, 947f.
[156] Ebd.

E-Person helfen würde, wenn diese kein Vermögen hat, auf das man zugreifen kann. Dieses Thema könnte bei zunehmender Autonomie von intelligenten Maschinen eine größere Rolle spielen. Zurzeit erscheint es jedoch eher fernliegend.[157]

b. Fondlösung

Nach der teilweise vorgeschlagenen Fondlösung soll von einer Haftung für durch KI-Systeme verursachte Schäden abgesehen werden.[158] Vielmehr soll ein staatlich kontrollierter Fond zur Entschädigung der Opfer eingerichtet werden.[159] Kritisiert wird diesbezüglich jedoch, dass dies dem einzelnen Schädiger die Motivation zur Schadensvermeidung nehmen könnte.[160] Die EU-Kommission hat in ihrem Bestreben, KI-Anwendung zu regulieren, jedoch letztlich auch einen anderen Weg gewählt.

2. Reformvorhaben der EU-Kommission

Um die rechtlichen Fragen, die sich im Zusammenhang mit der Nutzung von KI stellen, zu regeln, hat die EU-Kommission in den letzten Jahren zwei Gesetzgebungsverfahren auf den Weg gebracht, die jedoch bislang noch nicht beschlossen wurden. Zum einen handelt es sich um den Entwurf einer Verordnung zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz 2021/0106/COD (im Weiteren *KI-Verordnung*), zum anderen um den Ende September 2022 vorgelegten Richtlinienvorschlag zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz 2022/0303/COD (im Weiteren: *Richtlinie über KI-Haftung*).

Die Richtlinie und die Verordnung sind im Zusammenhang zu sehen und beziehen sich in wesentlichen Teilen aufeinander. Für die sich im Rahmen dieses Gutachtens stellenden Haftungsfragen haben die Regelungen des Richtlinienvorschlags jedoch die größere Bedeutung. Diese sind in wesentlichen Teilen schwächer ausgestaltet, als ursprünglich zu erwarten gewesen war. So wurde in einer Entschließung des Europäischen Parlaments aus dem Jahre 2020 noch eine verschuldensunabhängige Betreiberhaftung gefordert.[161] Eine solche sieht der Richtlinienvorschlag jedoch explizit (noch nicht) vor.[162]

---

[157] *Mühlböck/Taupitz*, AcP 221 (2021), 179, 216.
[158] Mühlböck/Taupitz, AcP 221 (2021), 179, 213; *Brand*, MedR 2019, 943, 947.
[159] Ebd.
[160] Ebd.
[161] *Europäisches Parlament*, Entschließung für eine Regelung der zivilrechtlichen Haftung beim Einsatz künstlicher Intelligenz (2020/2014(INL)), Art 4.
[162] *Europäische Kommission*, Richtlinienvorschlag außervertragliche zivilrechtliche Haftung an künstliche Intelligenz 2022/0303/COD, S. 10.

Vielmehr wird ein stufenweiser Ansatz vorgeschlagen.[163] So sollen zunächst Maßnahmen *zur Erleichterung der Beweislast für Opfer, die ihre Haftungsansprüche nachweisen wollen*, eingeführt werden.[164] Artikel 5 der *Richtlinie über KI-Haftung* legt ein Überprüfungsdatum fest, bis zu welchem von der EU-Kommission evaluiert werden soll, ob die getroffenen Maßnahmen ausreichen, um die Ziele der Richtlinie (das Schließen von etwaigen Haftungslücken bei Einsatz von KI[165]) zu erreichen oder ob insbesondere die Einführung einer Gefährdungshaftung geboten ist. Eine KI-Betreiberhaftung in Form einer Gefährdungshaftung scheint daher zunächst nicht zu kommen. Allerdings wird aus der Regelung auch klar, dass sie rechtspolitisch noch nicht verworfen wurde.[166]

Was die *Richtlinie über KI-Haftung* der EU-Kommission schon jetzt vorsieht, sind Neuregelungen im Hinblick auf das Beweisrecht. Diese finden sich insbesondere in Art. 3 und 4. Nach Art. 1 betreffen die Regelungen der Richtlinie jedoch nur außervertragliche Schadensersatzansprüche und damit das deutsche Deliktsrecht, jedoch nicht Ansprüche aus Behandlungsvertrag. Allerdings ist es – aufgrund der engen Verzahnung von deliktischen und vertraglichen Ansprüchen im Arzthaftungsrecht nur schwer vorstellbar, dass Beweiserleichterungen im außervertraglichen Bereich, nicht auch Auswirkungen auf vertragliche Ansprüche im Arzthaftungsprozess haben werden. Nach Art. 3 Abs. 1 sollen Regelungen geschaffen werden, die es einem Kläger ermöglichen, entweder von einem Nutzer oder einem Anbieter eines Hochrisiko-KI-Systems, das im Verdacht steht, einen Schaden verursacht zu haben, die einschlägigen Beweismittel herauszuverlangen. Sowohl im Hinblick auf den Begriff des „Nutzers", als auch auf den Begriff des „Hochrisiko-KI-Systems" wird über Art. 2 auf die *KI-Verordnung* verwiesen. Nach Art. 3 Nr. 4 der *KI-Verordnung* können Ärzte bzw. Krankenhäuser „Nutzer" sein. KI-Software, die für medizinische Zwecke genutzt wird, dürfte in aller Regel auch als „Hochrisiko-KI-System" einzustufen sein.[167] Art. 6 der *KI-Verordnung*, der die Voraussetzungen für KI-Systeme benennt, verweist diesbezüglich auf den Anhang II. Dort wird in Nr. 1 auf die MedProdVO (EU) 2017/ 745 verwiesen, welche – wie bereits oben festgestellt – für Medizinprodukte einschlägig ist.[168]

Für KI-Software, die im medizinischen Bereich eingesetzt wird, ist Art. 3 der *KI-Haftungsrichtlinie* mithin einschlägig. Der Umfang der Offenlegungspflicht wird in Art. 3 Abs. 4 bestimmt. Danach beschränken die Gerichte *die Offenlegung von Beweismitteln und die Maßnahmen zu deren Sicherung auf das Maß, das erforderlich und verhältnismäßig ist, um einen Schadensersatzanspruch eines Klägers oder potenziellen Klägers zu stützen.* Hier muss vor allen

[163] Ebd.
[164] Ebd.
[165] Ebd. S. 1ff.
[166] Siehe dazu auch: *Bomhard/Siglmüller*, RDi 2022, 506.
[167] Siehe hierzu auch: *Zech/Hünefeld*, MedR 2023, 1, 6ff.
[168] Siehe C.I.2.a.

Dingen das Interesse am Schutz von Betriebsgeheimnissen mit dem Interesse an der Durchsetzung eines Schadensersatzanspruches abgewogen werden, was in der Praxis zu großen Konflikten zwischen den Parteien führen könnte.[169] Dies gilt umso mehr vor dem Hintergrund, dass Art. 3 Abs. 5 des Richtlinienvorschlags die Vermutung eines Sorgfaltspflichtverstoßes vorsieht, wenn ein Beklagter der Offenlegung von Beweismitteln nicht nachkommt.

Art. 4 der *Richtlinie über KI-Haftung* konstatiert unter bestimmten Bedingungen (Nachweis des Verschuldens, Einfluss des Verschuldens auf Ergebnis der KI, Nachweis, dass das KI-Ergebnis zu einem Schaden geführt hat) die Vermutung eines ursächlichen Zusammenhangs zwischen dem Verschulden des Beklagten und dem vom KI-System hervorgebrachten Ergebnis. Ein Sorgfaltspflichtverstoß und damit einhergehend ein Verschulden des Beklagten muss jedoch weiterhin vom Kläger nachgewiesen werden.[170] Für den Nachweis des Verschuldens statuiert Art. 4 Abs. 3 erweiterte Nachweispflichten des Klägers gegenüber dem beklagten Nutzer. So muss nachgewiesen werden, dass der Nutzer:

Seiner Pflicht zur Verwendung oder Überwachung des KI-Systems entsprechend der beigefügten Gebrauchsanweisung oder gegebenenfalls zur Aussetzung oder Unterbrechung seiner Verwendung […] nicht nachgekommen ist oder

Eingabedaten, die seiner Kontrolle unterliegen, auf das KI-System angewandt hat, die der Zweckbestimmung des Systems […] nicht entsprechen.

Inwieweit sowohl die *KI-Verordnung* als auch die *Richtlinie über KI-Haftung* nach Durchlaufen der parlamentarischen Verfahren noch Änderungen unterworfen sein werden, lässt sich derzeit nicht absehen. Festzuhalten bleibt jedoch, dass es eine Gefährdungshaftung für Nutzer von KI-Software vorerst nicht geben wird.[171] Inwiefern die in Art. 3 und 4 der Richtlinie über KI-Haftung getroffenen Beweisregelungen größere Veränderungen in Bezug auf die Nutzerhaftung bringen werden, bleibt abzuwarten. Gerade die in Art. 4 Abs. 3 formulierten erweiterten Nachweispflichten des Klägers, um in den Genuss der Vermutungswirkung des Art. 4 Abs. 1 zu kommen, könnten dazu führen, dass die Regelung in Arzthaftungsprozessen eher geringe Auswirkungen haben dürfte. Hier ist jedoch die konkrete Ausgestaltung abzuwarten.

[169] *Bomhard/Siglmüller*, RDi 2022, 506, 508f.
[170] *Bomhard/Siglmüller*, RDi 2022, 506, 510f.
[171] Siehe hierzu auch: Wagner, JZ 2023, 123, 133.

D. Zusammenfassung der Ergebnisse

- Die Nutzung von KI-Software im medizinischen Bereich stellt nicht per se eine Pflichtverletzung dar.[172]
- KI-Software hat das Potenzial, in Zukunft den ärztlichen Standard zu prägen.[173] Zurzeit ist die Nutzung von KI-Software zur Behandlung und Diagnostik jedoch als *Neulandmethode* einzustufen.[174]
- Sollte KI-Software in Zukunft den ärztlichen Standard prägen, so würde sich aller Voraussicht nach auch die Standardbestimmung, weg von Empfehlungen von Behandlungsmethoden von Krankheitsbildern, hin zu Handlungsempfehlungen bezogen auf bestimmte Patientengruppen, entwickeln.[175] Eine Nichtnutzung von KI-Software, die zum Standard geworden ist, dürfte erst nach einer Übergangsphase geeignet sein, um zu einer Pflichtverletzung zu führen.[176]
- Die Nutzung von KI-Software hat de lege lata keinen großen Einfluss auf die Arzthaftung.[177] Insbesondere werden Maschinenfehler nicht über § 278 BGB zugerechnet.[178] Auch eine Haftung über § 833 BGB ist abzulehnen.[179]
- Im Hinblick auf das „Intelligenzrisiko“[180] bestehen besondere Aufklärungspflichten gegenüber dem Patienten nach § 630e BGB. Dieser sollte – auch wenn die Nutzung von KI-Software zum Standard geworden sein sollte – weiterhin über das Intelligenzrisiko aufgeklärt werden.[181]
- Die Beweislastumkehr nach § 630h Abs. 1 BGB greift bei Schäden durch den Einsatz von KI-Software nicht, da KI-Software aufgrund des „Intelligenzrisikos“[182] für den Behandelnden nicht voll beherrschbar im Sinne des § 630h Abs. 1 BGB ist.[183]
- Es bleibt abzuwarten, wie sich die Arzthaftung bei der Nutzung von KI-Software im Hinblick auf die Reformvorhaben der EU-Kommission entwickeln wird. Nach dem aktuellen Richtlinienvorschlag der EU wird eine Gefährdungshaftung für den Betreiber von KI-Software zunächst nicht kommen.[184]

---

[172] Siehe C.I.1.
[173] Siehe C.I.2.b.
[174] Siehe C.I.2.a.
[175] Siehe C.I.2.b.
[176] Ebd.
[177] Siehe C.I.4.; C.II.5.
[178] Siehe C.I.3.b.
[179] Siehe C.II.3.
[180] Siehe A.III.3.
[181] Siehe C.III.
[182] Siehe A.III.3.
[183] Siehe C.IV.
[184] Siehe C.V.2.

Anlage 4: Fragebogen zur Nutzerbefragung.

Sehr geehrte Patientin, sehr geehrter Patient,

Sie leiden unter einer **peripheren arteriellen Verschlusskrankheit (PAVK)**, einer häufigen Volkskrankheit, die sich in früheren Stadien durch Beinschmerzen unter Belastung (sog. „Schaufensterkrankheit“) und in fortgeschrittenen Stadien auch mit ischämischen Ruheschmerzen ohne Belastung sowie Wundheilungsstörungen (sog. „Ulcus cruris“) äußern kann.
Mit dieser anonymen Umfrage möchten wir herausfinden, ob **moderne digitale Medien** geeignet sind, den Krankheitsverlauf bei Patienten/-innen mit einer PAVK günstig zu beeinflussen.

Geschlecht ○ männlich ○ weiblich ○ divers

Alter (Jahre) _____ **PAVK bekannt seit** (Jahre) _____

Chronische kritische Ischämie oder diabetisches Fußsyndrom:
Jemals **Beinschmerzen in Ruhe** oder **durchblutungsbedingte Wundheilungsstörungen** (Ulcera, Gangrän, Nekrose)? ○ JA ○ NEIN

Höchster Bildungsabschluss
○ Keine Angaben ○ Ohne Abschluss ○ Lehre/Berufsausbildung
○ Fachschulabschluss ○ Bachelor ○ Master (o. vergleichb.)
○ Diplom (o. vergleichb.) ○ Promotion

Wohnregion ○ städtisch ○ intermediär ○ ländlich

Wie informieren Sie sich über Gesundheitsthemen (Mehrfachauswahl)
☐ TV ☐ Zeitung ☐ Internet ☐ Radio ☐ Andere

	Bitte wählen Sie	
Seit der PAVK-Erstdiagnose, habe ich mein Verhalten geändert	○ JA	○ NEIN
Ich besitze (oder habe Zugriff auf) ein Smartphone	○ JA	○ NEIN
Wenn ja: Ich nutze damit auch mobile Applikationen (sog. „Apps“)	○ JA	○ NEIN
Ich habe selbst auch schon mal „Gesundheits-Apps“ verwendet	○ JA	○ NEIN
Ich benötige regelmäßig Unterstützung bei der Handynutzung	○ JA	○ NEIN
Ich besitze eine Pulsuhr, Schrittzähler oder andere „Wearables“	○ JA	○ NEIN

Gesundheits-Apps, die mein Verhalten beobachten & auswerten, könnten mir helfen, gesünder zu leben	○ JA	○ VLLT	○ NEIN
Ich kenne den Grund für jedes Medikament, das mir derzeit verschrieben wird	○ JA	○ VLLT	○ NEIN

1: Stimme ganz und gar nicht zu, 2: Stimme nicht zu, 3: Stimme zu, 4: Stimme voll und ganz zu

Die Bestimmung des individuellen Gesundheitsrisikos über Online-Kalkulatoren, wie score.germanvasc.de, ist nützlich	○ 1	○ 2	3	○ 4
Ich fand die Darstellung und Beschreibung der Inhalte unter score.germanvasc.de verständlich	○ 1	○ 2	3	○ 4
Ich würde Online-Kalkulatoren, wie score.germanvasc.de, zur Ergänzung der ärztlichen Beratung verwenden	○ 1	○ 2	3	○ 4

pAVK - Fragen zur Ernährung

1. Haben Sie in Verbindung mit Ihrer Erkrankung Informationen zur gesunden Ernährung erhalten?
O nein
O ja, schriftliche Informationen (Faltblatt, Broschüre)
O ja, ärztliche Hinweise
O ja, eine ausführliche Beratung durch eine Ernährungsfachkraft
O keine Angabe/weiß ich nicht

Bitte denken Sie bei der Beantwortung der folgenden Fragen zu Ihren Ernährungsgewohnheiten und zu Getränken an die letzten 12 Monate. Berücksichtigen Sie dabei auch Lebensmittel und Getränke, die Sie außerhalb des Hauses (Kantine, Restaurant etc.) konsumieren:

2. Wie häufig essen Sie Obst (z.B. 1 Apfel, 1 Banane, 1 Handvoll Erdbeeren, 1 kleiner Obstsalat)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

3. Wie häufig verzehren Sie Hülsenfrüchte (z.B. Linsen, Erbsen, Bohnen als Eintopf oder Beilage)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

4. Wie häufig essen Sie Gemüse, Salat oder Rohkost (z.B. als Beilage zum Hauptgericht oder Salatteller)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

5. Wie häufig verzehren Sie Fisch (z.B. als Hauptgericht, Fischbrötchen)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche

O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

6. Wie häufig verzehren Sie rotes Fleisch (Rind, Schwein, Schaf, Lamm) als Hauptgericht und daraus hergestellte Wurstwaren z.B. als Aufschnitt?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

7. Wie häufig verzehren Sie Geflügel (Huhn, Pute) als Hauptgericht und daraus hergestellte Wurstwaren z.B. als Aufschnitt
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

8. Wie häufig essen Sie Vollkornbrot, Getreideflocken, Vollkornreis, -nudeln und Kartoffeln (z.B. 1 Scheibe Brot, 2-3 Esslöffel Müsli, Reis, Nudeln als Beilage, 2 mittelgroße Kartoffeln?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

9. Wie häufig verzehren Sie Milch und Milchprodukte wie Joghurt, Quark, Käse (z.B. 1 Glas Milch, 1 kleiner Becher Joghurt, 1 Scheibe Käse)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

10. Bevorzugen Sie in der Regel <u>fettreduzierte</u> Milch und Milchprodukte?

O ja
O nein
O keine Angabe/weiß ich nicht

11. Wie häufig verwenden Sie Butter und Sahne (z.B. 2 Teelöffel Butter, 1-2 Esslöffel Sahne)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

12. Wie häufig verwenden Sie pflanzliche Öle wie Oliven-, Raps- oder Sonnenblumenöl (z.B. 1 Esslöffel zur Zubereitung von Salat, Gemüse, gekochten Speisen)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

13. Welche pflanzlichen Öle verwenden Sie überwiegend zur Zubereitung von Salat, Gemüse und gekochten Speisen?
O Rapsöl
O Olivenöl
O andere pflanzliche Öle
O keine Angabe/weiß ich nicht

14. Wie häufig verzehren Sie <u>ungesalzene</u> Nüsse wie Walnüsse, Haselnüsse, Mandeln (z.B. 1 kleine Handvoll)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

15. Wie häufig verzehren Sie Kekse, Kuchen und Süßigkeiten (z.B. 1 Stück Kuchen, 2-3 Kekse, 1 Riegel Schokolade)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich

O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

16. Wie häufig verzehren Sie <u>gesalzene</u> Snacks wie Chips, Erdnüsse, Salzstangen (z.B. 1 kleine Handvoll)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

17. Wie häufig trinken Sie gesüßte Getränke wie Limonade, Cola, Eistee oder Säfte mit zugesetztem Zucker (z.B. Fruchtnektar, Fruchtsaftgetränke)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

18. Wie häufig konsumieren Sie alkoholische Getränke wie Wein, Sekt, Bier oder Schnaps (z.B. 1 Glas Wein, 1 kleine Flasche Bier (0,3 l), 1 kleines Glas Schnaps)?
O selten oder gar nicht
O 1 x pro Woche
O 2 x pro Woche
O 3 x pro Woche
O 4-6 x pro Woche
O 1 x täglich
O 2 x täglich
O 3 x täglich oder mehr
O keine Angabe/weiß ich nicht

19. Nehmen Sie regelmäßig Vitamin- oder Mineralstoffpräparate ein?
O nein
O ja, Vitaminpräparate
O ja, Mineralstoffpräparate
O ja, Vitamin- und Mineralstoffpräparate
O keine Angabe/weiß ich nicht