DGHO
DEUTSCHE GESELLSCHAFT FÜR
HÄMATOLOGIE UND MEDIZINISCHE ONKOLOGIE

GESUNDHEITSPOLITISCHE SCHRIFTENREIHE DER DGHO

Band 21

Potenzial und Barrieren bei der Nutzung von Patientendaten in KI-basierter Forschung am Beispiel der Hämatologie und Medizinischen Onkologie

Plädoyer für eine verantwortungsvolle und pragmatische Ressourcennutzung

Potenzial und Barrieren bei der Nutzung von Patientendaten in KI-basierter Forschung am Beispiel der Hämatologie und Medizinischen Onkologie

Plädoyer für eine verantwortungsvolle und pragmatische Ressourcennutzung

Gesundheitspolitische Schriftenreihe der DGHO
Band 21

Potenzial und Barrieren bei der Nutzung von Patientendaten in KI-basierter Forschung am Beispiel der Hämatologie und Medizinischen Onkologie

Stand: Mai 2024
ISBN: 978-3-9821204-5-4

Herausgeberinnen und Herausgeber:
Prof. Dr. med. Andreas Hochhaus
Prof. Dr. med. Claudia Baldus
Prof. Dr. med. Martin Bentz
Dr. med. Carsten Oliver Schulz

DGHO Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e. V.

www.dgho.de
info@dgho.de

Autorinnen und Autoren:

Christian Pohlkamp*, Carsten Marr, Jan Moritz Middeke*****
Arbeitskreis KI der Dt. Gesellschaft für Hämatologie und Onkologie,
*Münchener Leukämielabor, **Helmholtz Munich, ***Universitaetsklinikum Dresden

Wiebke Rösler
Universitätsspital Zürich, Walter-Siegenthaler-Gesellschaft, Arbeitskreis KI der Dt. Gesellschaft für Hämatologie und Onkologie

Jakob Nikolas Kather
Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Arbeitskreis KI der Dt. Gesellschaft für Hämatologie und Onkologie

Michael Heuser
Klinik für Hämatologie, Hämostaseologie, Onkologie und Stammzelltransplantation für Hämatologie, Medizinische Hochschule Hannover

Melanie Börries
Institut für Medizinische Bioinformatik und Systemmedizin, Universitätsklinikum Freiburg

Christian Thielscher
KompetenzCentrum für Medizinoekonomie, FOM Hochschule für Oekonomie & Management

Bernd Eylert
TH Wildau (Technical University of Applied Sciences Wildau)

Felix Nensa
Universitätsklinikum Essen (AöR), Institut für Diagnostische und Interventionelle Radiologie und Neuroradiologie, Institut für Künstliche Intelligenz in der Medizin (IKIM), Smart Hospital Information Platform (SHIP) Data Integration Center (DIC)

Rickmer Braren
Institut für Radiologie, TU München

Philipp Müller-Peltzer
Kanzlei Schürmann/Rosenthal/Dreyer, Rechtsanwälte für Digitales Business, Technologie & Medien

Christoph Kornauth
Münchner Leukämielabor

Christopher Braun[ab], Lena Lörcher[a], Marco Huber[a,b]
[a]Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, Stuttgart, Deutschland
[b]Institut für Industrielle Fertigung und Fabrikbetrieb IFF, Universität Stuttgart, Deutschland

Satz:
PRINTCOUTURE®, Tinajo

INHALTSVERZEICHNIS

Vorwort des DGHO-Vorstands

Liebe Leserinnen und Leser,
liebe Kolleginnen und Kollegen,

Im Jahr 2013 wurde der 1. Band der Gesundheitspolitischen Schriftenrehe der DGHO „Herausforderung demografischer Wandel. Bestandsaufnahme und künftige Anforderungen an die onkologische Versorgung“ veröffentlicht. Ziel der Gesundheitspolitischen Schriftenreihe war und ist es, medizinische, gesundheitspolitische und ethischen Themen aufzugreifen, die uns im Rahmen unserer wissenschaftlichen und ärztlichen Tätigkeit begegnen.

Dass wir uns in der Hämatologie und Medizinischen Onkologie in einem breiten Spektrum bewegen, macht die thematische Breite der bisherigen Bände deutlich: Arzneimittelengpässe, ärztlich assistierte Selbsttötung, Demographischer Wandel, Forschung, Förderung von Ärztinnen, Frühe Nutzenbewertung, junge Erwachsene mit Krebs, Krebsfrüherkennung, Medizin und Ökonomie, Pflege.

Kaum mehr als eine Dekade nach der Publikation des 1. Bandes hat der DGHO-Arbeitskreis „Künstliche Intelligenz in der Hämatologie und Onkologie“ ein umfangreiches Positionspapier erarbeitet, das nun in Form des vorliegenden 21. Bandes der Gesundheitspolitischen Schriftenreihe veröffentlicht worden ist. Inmitten der aktuell intensiv geführten Debatte zum Thema der Gesundheitsdatennutzung skizziert der Arbeitskreis – in Zusammenarbeit mit externen Expertinnen und Experten – das Potenzial der auf Künstlicher Intelligenz (KI) basierenden Gesundheitsdatenforschung und stellt diesem strukturelle Mängel und rechtliche Einschränkungen gegenüber.

Thematische Schwerpunkte sind unter anderem der Einsatz von Künstlicher Intelligenz für den medizinischen Fortschritt und die Verbesserung der medizinischen Versorgung in der Bundesrepublik Deutschland. In diesem Zusammenhang werden die aktuell bestehenden Defizite im Bereich der Datenspeicherung beispielsweise in Folge einer Vielzahl parallel genutzter IT-Systemen dargestellt.

Darüber hinaus wird die Rolle datenschutzrechtlicher Barrieren und Unsicherheiten bei der Rechtsauslegung beleuchtet. Schließlich werden konkrete Handlungsempfehlungen abgeleitet.

Laut ChatGPT 4 „besitzt die Künstliche Intelligenz das Potenzial für erhebliche wirtschaftliche Veränderungen, birgt soziale, ethische, philosophische und existenzielle Herausforderungen, führt zu einer Erweiterung menschlicher Fähigkeiten und bedarf einer internationalen Regulierung".

Die – auf ChatGPT basierende – Aussage „Es ist nicht die Frage, ob Künstliche Intelligenz fundamentale Veränderungen mit sich bringen wird. Vielmehr wird die Art und Weise der Integration und Nutzung der Technologie dafür entscheidend sein, wie sie die menschliche Entwicklung in Zukunft prägen wird" wirkt fast schon wie ein Allgemeinplatz. Nun zeichnen sich Allgemeinplätze dadurch aus, dass sie sich nicht ohne Grund im kollektiven Sprachgebrauch etabliert haben. Bei der Künstlichen Intelligenz erleben wir dieses Phänomen allerdings in einem bisher nie dagewesenen Tempo.

Wir danken dem DGHO-Arbeitskreis „Künstliche Intelligenz in der Hämatologie und Onkologie" für die Erarbeitung des vorliegenden Beitrags. Die Publikation macht deutlich, wie durch einen fachbereichs-übergreifenden Dialog in kollaborativem Geist auch komplexe Themen proaktiv und zielführend ausgestaltet werden können.

Mit herzlichen kollegialen Grüßen

Prof. Dr. med. Andreas Hochhaus
Geschäftsführender Vorsitzender

Prof. Dr. med. Claudia Baldus
Vorsitzende

Prof. Dr. med. Martin Bentz
Mitglied im Vorstand

Dr. med. Carsten-Oliver Schulz
Mitglied im Vorstand

1

ZUSAMMENFASSUNG

1. Zusammenfassung

Die moderne medizinische Diagnostik und Therapie erzeugen eine Vielzahl digitaler Daten. Die Gesamtheit der erfassbaren und bereits erfassten Gesundheitsdaten stellt ein enormes Potenzial für die medizinische Forschung und insbesondere für den Einsatz Künstlicher Intelligenz (KI) dar. KI-unterstützte Forschung kann eine umfassende Verbesserung der medizinischen Versorgung in Deutschland ermöglichen. Dieses Potenzial wird aktuell aufgrund begrenzter Datenverfügbarkeit nur unzureichend und weniger als in vielen anderen Ländern genutzt. Hierfür existieren verschiedene Gründe: Die Datenspeicherung erfolgt in Deutschland infolge einer Vielzahl diagnostisch und therapeutisch tätiger Institutionen und genutzter IT-Systeme quantitativ, qualitativ und strukturell inhomogen. Die Auswertung unstrukturierter Daten mittels großer Sprachmodelle (engl. Large Language Models, LLMs) schafft hier nur in Teilen Abhilfe, denn vielfach mangelt es an Schnittstellen zur Datenübermittlung zwischen verschiedenen Plattformen. Grundsätzlich fehlen zudem klare Anreize für eine systematischere primäre Dokumentation der Daten, insbesondere auf Ebene medizinischer Leistungserbringer. Eine erhebliche Rolle spielen auch datenschutzrechtliche Barrieren und Unsicherheiten bei der Rechtsauslegung.

Eine systematische Digitalisierung, Anonymisierung und Zusammenführung diagnostischer und therapeutischer Daten kann deren Nutzen um ein Vielfaches erhöhen. Technische Voraussetzung dafür ist neben der Datenerhebung die Etablierung einer leistungsfähigen, interoperablen, digitalen Infrastruktur. Rechtlich hingegen ist ein pragmatischer Standard für den Umgang mit bestehenden Datenschutzbestimmungen und für deren konkrete Umsetzung im Rahmen der aktuellen Gesetzeslage dringend erforderlich. In Teilen erscheint eine kritische Revision bestehender Gesetze angebracht, da aktuelle technologische Möglichkeiten nicht immer angemessen berücksichtigt werden.

Ziel dieses Positionspapiers ist es, das Potenzial der KI-basierten Gesundheitsdatenforschung anhand von Beispielen zu skizzieren und den strukturellen Mängeln und rechtlichen Einschränkungen gegenüberzustellen. Stellvertretend wird der Bereich der Hämatologie und Onkologie behandelt, der aufgrund der Heterogenität der Erkrankungen, der hohen Innovationsdynamik und des starken sozioökonomischen Einflusses als Beispiel besonders geeignet ist. Initiiert von den MedizinerInnen und Medizinforschenden des Arbeitskreises KI der Deutschen Gesellschaft für Hämatologie und Medizinische Onkologie (DGHO) und getrieben vom Wunsch, Patienten bestmögliche Behandlung zu bieten sowie von der Besorgnis über ein Zurückfallen des Forschungs- und Gesundheitsstandortes Deutschland im internationalen Vergleich wurde in Zusammenarbeit mit ExpertInnen und InteressenvertreterInnen eine gemeinsame Position erarbeitet und mit ähnlich ausgerichteten Initiativen abgestimmt.

Die parallel zur Ausarbeitung dieses Konsenses erfolgte Verabschiedung des Digitalgesetzes und des Gesundheitsdatennutzungsgesetzes (GDNG) im Dezember 2023 greift viele der oben genannten Aspekte auf. Hinzu kommen das Medizinforschungsgesetz, das Digitalagenturgesetz und die europäischen Vereinbarungen zum European Health Data Space (EHDS). Die (teils noch vorläufigen) Entwürfe stoßen aber aufgrund diverser planerischer Unwägbarkeiten auf Bedenken, sowohl bei Teilen der Gesundheitsversorgung als auch bei den Datenschutzinstitutionen.

Die AutorInnen und UnterstützerInnen dieses Positionspapiers mahnen eine Anpassung der oben genannten Rahmenbedingungen an den rasanten technologischen Fortschritt an. Insbesondere wird gefordert:

1. Eine digitale Dokumentation von Gesundheitsdaten, samt einer Verpflichtung zum Angebot standardisierter Schnittstellen durch IT-Hersteller
2. Adäquate Anreize und Personalressourcen für die Erhebung und Bereitstellung von Gesundheitsdaten, auch auf der Ebene von medizinischen Primärversorgern
3. Eine leistungsstarke und sichere technische Infrastruktur für die Verarbeitung der großen Mengen an Gesundheitsdaten
4. Eine pragmatische länderübergreifende Gestaltung und Auslegung der Einwilligungs- und Datenschutzbestimmungen, die das medizinische Gemeinwohl und die Heilung schwerer Erkrankungen angemessen berücksichtigt, unnötige Hürden beim Datenzugang für die Forschung beseitigt und den großen Wert umfassender medizinischer Daten anerkennt.

Nur unter zügiger, staatlich geförderter und kontrollierter Umsetzung dieser Forderungen kann eine nachhaltige, datengetriebene und evidenzbasierte Verbesserung der deutschen Gesundheitsversorgung sichergestellt und der Wissenschafts- und Wirtschaftsstandort Deutschland wieder konkurrenzfähiger gemacht werden. Insbesondere auf die Gestaltung, praktische Auslegung und Akzeptanz der jüngst verabschiedeten oder kurz vor der Verabschiedung stehenden Gesetze (Digitalgesetz, Gesundheitsdatennutzungsgesetz , Medizinforschungsgesetz, Digitalagenturgesetz) wird es hier entscheidend ankommen.

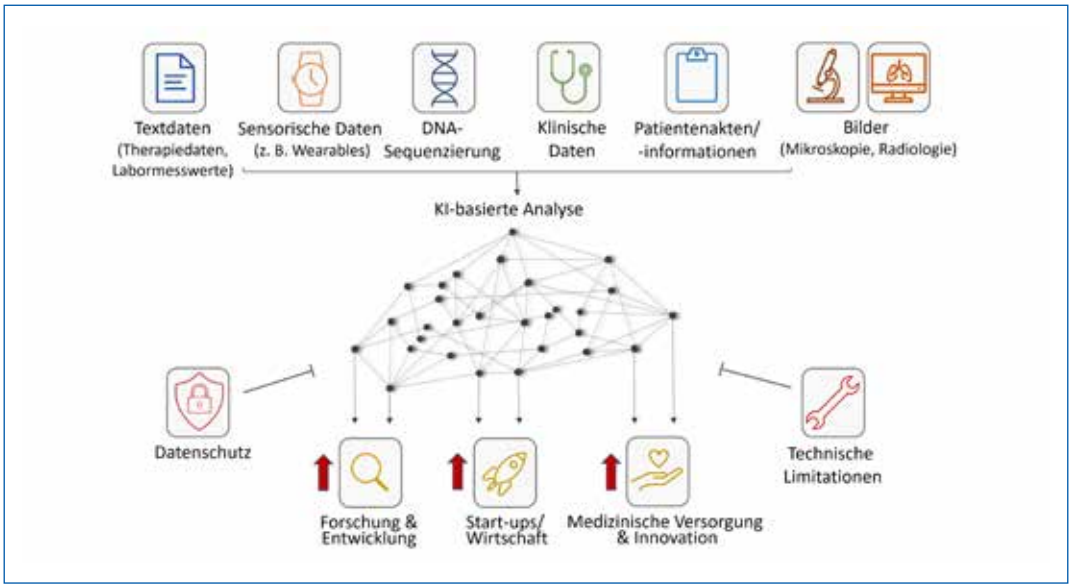
2

HINTERGRUND

2. Hintergrund

Unser Gesundheitssystem, das zurecht als eines der besten der Welt gilt, erzeugt täglich hochwertige und detaillierte Daten in großem Umfang und zunehmend in digitalisierter Form. Dabei kann es sich z.B. um Bilddaten (u.a. aus der Radiologie oder Mikroskopie), genetische Daten (z.B. DNA-Sequenzen, Mutationsdaten) oder Textdaten (z.B. Therapiedaten oder Labormesswerte) handeln (siehe Abb. 1). Gleichzeitig stehen diese Daten aber der institutionellen und kommerziellen Forschung in nur sehr begrenztem Maße zur Verfügung. Dies betrifft alle medizinischen Disziplinen, insbesondere auch die Hämatologie und Onkologie, welche auf eine datenbasierte Charakterisierung von Krebserkrankungen angewiesen sind. Nur so können effizientere Therapien und damit höhere Heilungschancen für die Vielzahl der betroffenen Patienten erreicht werden.

Abbildung 1: Potenziale und Barrieren bei der wissenschaftlichen Nutzung von Gesundheitsdaten



Grafik: MLL Münchner Leukämielabor

Für die fehlende Verfügbarkeit von Gesundheitsdaten in der Forschung gibt es mehrere Gründe: In den verschiedenen klinischen Einrichtungen (Krankenhäuser, Labore, Arztpraxen etc.) werden solche Daten – wenn überhaupt – in unterschiedlichem Detailgrad, unterschiedlicher Strukturierung und nach uneinheitlichen Prinzipien digital gespeichert. Dies ist einerseits auf die große Varianz der genutzten IT-Systeme und die fehlende Vereinheitlichung und Automatisierung zurückzuführen. Andererseits bestehen in Datenschutzfragen häufig zumindest Unsicherheiten, oft werden aber auch klare gesetzliche Hürden wahrgenommen: Die Interpretation der europäischen Datenschutzrichtlinie erfolgt auf Länderebene nicht immer einheitlich.

Schließlich fehlt es oft an Anreizen für die klinische Institution, im eng getakteten Arbeitsalltag überhaupt einen Beitrag zu einer systematischen Datenerhebung zu leisten, die über das Notwendigste hinausgeht.

An diesen Hindernissen gilt es anzusetzen, um das immense Potenzial der Gesundheitsdaten zu nutzen. Dies hat auch die Politik erkannt und ein ganzes Paket von Gesetzen angekündigt und zum Teil bereits verabschiedet, welches den Notwendigkeiten Rechnung tragen soll. Das Gesundheitsdatennutzungsgesetz, das Digitalgesetz, das Medizinforschungsgesetz und das Digitalagenturgesetz sollen unter anderem den Zugang zu deutschen Gesundheitsdaten für die Forschung erleichtern. Auch auf europäischer Ebene gibt es weit fortgeschrittene Projekte, wie z.B. den geplanten EHDS (European Health Data Space), bei dem im März 2024 in buchstäblich letzter Minute eine Einigung über die Ausgestaltung der Opt-out-Regelung und einiger anderer Kernelemente erzielt wurde.

Um entsprechende Chancen zu realisieren, sind aufgrund der riesigen zu prozessierenden Daten, Anwendungen und Modelle aus dem Feld der Künstlichen Intelligenz (KI) essenziell. Diese sind in der Lage, in großen Datenmengen Muster und Zusammenhänge zu erkennen und damit valide Schlussfolgerungen zu ziehen. Erfreulicherweise findet sich auch in der politischen und gesellschaftlichen Diskussion gegenüber KI-basierten Methoden – bei allen berechtigten Bedenken – eine zunehmende Offenheit, nicht zuletzt befeuert durch den Einzug entsprechender Technologien auch in andere Gesellschaftsbereiche.

Voraussetzung für die Anwendung solcher KI-Modelle ist neben der Etablierung entsprechender digitaler Infrastrukturen (inklusive hochpotenter Cloud-basierter Plattformen für multi-institutionelle Datenerfassung und Modelltraining) ein konsensfähiger Standard für die Datenerhebung und -speicherung sowie für den Umgang mit bestehenden Datenschutzbestimmungen. Insbesondere die Nutzung genetischer Daten wird datenschutzrechtlich kontrovers gesehen, birgt aber gleichzeitig das vermutlich höchste Potenzial. Hier stellt sich die Frage, inwiefern bezüglich der Dateninhaberschaft und Datenverwendung die gesamtgesellschaftliche Solidarität und Verantwortung des Individuums stärker in den Vordergrund rücken sollte, sowie welche objektivierbaren datenschutzrechtlichen Gefahren bestehen, denen durch eine konsequente Pseudonymisierung oder Anonymisierung nicht zur Genüge entgegengewirkt werden kann. Kann man in einer Kosten-Nutzen-Rechnung den Schaden, den das Zurückhalten von Daten künftigen Patientengenerationen – und in finanzieller Hinsicht auch dem deutschen Gesundheitssystem – zufügen wird, sinnvoll quantifizieren? Auch das Spannungsfeld zwischen institutioneller und kommerzieller Forschung an Gesundheitsdaten ist zu diskutieren, insbesondere, wenn das definierte Ziel die breite Verfügbarkeit KI-basierter diagnostischer Modelle ist, was eine Beteiligung kommerzieller Akteure voraus-

setzt. Die Auswirkungen auf den Wirtschaftsstandort Deutschland in einem äußerst zukunftsträchtigen Sektor sind ebenfalls von hoher Relevanz.

Die Idee zu diesem Positionspapier entstand im Rahmen des „Arbeitskreises KI" der Deutschen Gesellschaft für Hämatologie und Medizinische Onkologie (DGHO), in dem sich InteressenvertreterInnen aus Medizin, Forschung, Naturwissenschaft, Informatik und Datenwissenschaften zusammengeschlossen haben. In Anbetracht täglich wahrgenommener Einschränkungen in der Forschungsarbeit und einer inzwischen auch auf höchster politischer Ebene verstärkten Diskussion über die Zukunft des Forschungsstandorts Deutschland wurde die Notwendigkeit einer Stellungnahme aus dem Kreis der MedizinerInnen und Forschenden selbst gesehen. Für ausgewählte Themen wurden Beiträge externer Experten und „Key Opinion Leader" eingeholt.

Ziel dieses Positionspapiers ist einerseits eine für nicht-medizinische EntscheidungsträgerInnen und Stakeholder verständliche Darstellung des medizinischen und ökonomischen Potenzials von Gesundheitsdaten anhand unterschiedlicher Datentypen. Andererseits soll ein Beitrag zur Schaffung optimierter Standards im Umgang mit solchen Daten geleistet werden. Die von der Bundesregierung auf den Weg gebrachte Gesetzgebung stellt ebenso wie Initiativen auf EU-Ebene einen prinzipiell äußerst begrüßenswerten Schritt dar. Jedoch bleiben der genaue Wortlaut bzw. die konkrete Interpretation und Umsetzung der Gesetze in vielen Details abzuwarten. Zudem regt sich bereits jetzt Widerstand in Teilen der ärztlichen Versorgungslandschaft und des Datenschutzes. Wir plädieren dringend dafür, geeignete Maßnahmen – auch im Sinne unserer abschließend aufgeführten Empfehlungen – in der Praxis zeitnah und verpflichtend voranzutreiben, um ein dramatisches Zurückfallen Deutschlands in der Gesundheitsforschung im internationalen Vergleich zu verhindern.

Explizit soll der Schulterschluss mit ähnlich ausgerichteten Initiativen realisiert werden. So setzen sich u.a. das „Deutsche Konsortium für Translationale Krebsforschung" (DKTK), das „nationale Netzwerk genomische Medizin" (nNGM) und das „Nationale Centrum für Tumorerkrankungen" (NCT) seit Jahren für die Umsetzung einer forschungsorientierten personalisierten Medizin durch Nutzung diagnostischer Real-World-Daten ein. Das „Deutsche Netzwerk für personalisierte Medizin" (DNPM) strebt u.a. die Verlaufsdokumentation moderner zielgerichteter Therapien in einer virtuellen Plattform zur Nutzung für Machine Learning-basierte Forschungsansätze an. Der „Berufsverband niedergelassener Hämatologen und Onkologen" (BNHO) und die „Arbeitsgemeinschaft internistische Onkologie" (AIO) bieten mit „AIO-BNHO-CONNECT" eine Plattform für die Aggregation von Real-World-Daten aus dem gesamten Versorgungskosmos, einschließlich einer Genomdatenbank und einer virtuellen Biobank. Im „genomDE"-Konsortium sind mehr als ein Dutzend Initiativen aus Universitätsmedizin, Patientenvertretungen und IT gebündelt, und engagieren sich für die Nutzung genomischer Daten. Das im letzten Jahrzehnt präsentierte Konzept der „Wissen generie-

renden Versorgung“ (WGV) wurde von Interessensvertretern aus Politik, Krankenkassen, Medizin, Forschung und auch Patientenschaft entwickelt. Der BNHO beteiligt sich z.B. am Projekt „NeoWis“ (Netzwerk zur wissensgenerierenden Versorgung) und arbeitet an eigenen Hard- und Softwarelösungen für eine strukturierte Datenerhebung. Die altbekannte Problematik der bundesweit sehr uneinheitlichen Daten- und IT-Struktur wird auch über Förderprogramme der Bundesministerien für Gesundheit und Bildung und Forschung (siehe die „Nationale Dekade gegen Krebs“ NKD) adressiert. Datenschutzrechtliche Aspekte werden ebenfalls von etlichen Sachverständigengremien und Initiativen pragmatisch diskutiert, wie z.B. in Stellungnahmen des Zentrums für Medizinische Datennutzbarkeit und Translation (ZMDT) der Universität Bonn, des Wissenschaftsrats oder der Walter-Siegenthaler-Gesellschaft[1]. Ein wichtiger Akteur auf diesem Gebiet ist zudem die „Medizininformatikinitiative” (MII), die seit Jahren vom BMBF gefördert wird.

3

POTENZIAL

3. Potenzial

Moderne KI-Modelle haben das Potenzial, Krankheiten anhand genetischer Daten vorherzusagen[2,3], Zellen und Gewebe zu klassifizieren[4–7] und völlig neue Korrelationen zwischen Krankheit und möglicher Ursache herzustellen[8]. Zur Etablierung solcher Modelle ist die Forschung an sehr großen, repräsentativen und multi-institutionellen Patientendatensätzen unabdingbar. Im Rahmen eines derart verbesserten diagnostischen Verständnisses zeichnen sich deutliche Fortschritte für eine individualisierte Therapie („personalisierte Medizin") ab. Auch jenseits der Forschung können viele Einschränkungen menschlich ausgeführter Analysen (Konzentration und Ermüdung bei repetitiven Aufgaben, ungenügende Standardisierung, hoher Zeitaufwand) in der Routinediagnostik mit Hilfe von KI potenziell egalisiert werden. Dadurch können höhere Datenmengen in kürzerer Zeit prozessiert und die Sensitivität, Genauigkeit und Geschwindigkeit der Diagnosestellung sowie anderer Untersuchungsparameter verbessert werden.

Der Einsatz von Methoden, die in der Lage sind, inhaltliche Zusammenhänge aus großen, unstrukturierten Datenbeständen zu extrahieren, hat das Potenzial von KI-Modellen dramatisch erweitert. Für das Trainieren spezifischer, nachgelagerter Aufgaben, wie der Klassifikation von Krankheiten oder der Erkennung von Tumoren, können bei Bedarf bestehende, vortrainierte und frei zugängliche große Modelle (engl. Foundation Models) [19,20] verwendet werden. Mit diesen neuartigen KI-Modellen, wie z.B. großen Sprachmodellen (engl. Large Language Models, oder LLMs), können aus unstrukturierten Daten strukturierte Informationen abgeleitet werden[19,21], so dass in Zukunft auch unstrukturierte digitale Daten eine wichtige Ressource zur Wissensgenerierung darstellen können.

Ein Beispiel für geeignete Forschungsdatensätze sind diagnostische Genomdaten, die neben ihrem erheblichen Umfang auch durch die notwendige Kombination mit anderen diagnostischen Daten eine erhebliche Komplexität aufweisen. Derart große Datensätze bieten ein unermessliches Potenzial für die Entdeckung bisher unbekannter Zusammenhänge zwischen diagnostischen und klinischen Variablen, beispielsweise für die Prädiktion von Therapieansprechen oder -toxizität bei bestimmten Genotypen. Die vom Bundesministerium für Gesundheit initiierte nationale Genomstrategie (genomDE), die u.a. von zahlreichen renommierten universitären Zentren unterstützt wird, unterstreicht die herausragende Bedeutung genetischer Daten für die moderne datengetriebene und KI-gestützte Forschung. Genomische Daten stellen jedoch nicht nur die vermutlich wertvollste, sondern auch die datenschutzrechtlich kritischste Datenkategorie dar.

Ein spezifisches Potenzial bieten auch Bilddaten. Es handelt sich auch hier um meist große Datensätze (mit mehreren Gigabyte für MRT- oder Histologie-Bildern eines einzelnen Patienten), und auch hier ist eine Interpretation und Gewichtung aller Bilddetails durch potenziell „unvoreinge-

nommene“ KI-Modelle theoretisch überlegen, was sich bereits bei der Unterstützung menschlicher Diagnostiker bewährt hat (s.u.). Klar ist, dass derart riesige Datenmengen nicht von menschlichen Forschenden, sondern nur von adäquaten KI-Modellen als Ganzes prozessiert und ausgewertet werden können, wenngleich selbstredend eine menschliche Plausibilitätsprüfung unabdingbar bleibt. Dafür sind leistungsstarke Hardware-Strukturen wie Cloud-Computing-Systeme unerlässlich.

Große Bedeutung kommt auch dem Thema Datenvielfalt zu. Um ein realistisches Abbild der Realität zu erzeugen und die Robustheit von Modellen zu überprüfen, sind multi-institutionell (von verschiedenen Kliniken, Arztpraxen, Laboren, Pharmastudien, etc.) erhobene Daten in „KI-lesbarer” Struktur zwingend erforderlich. Eine inhaltlich homogene Datenerhebung wäre wünschenswert. Zudem sollten Daten nicht nur zu einem einzelnen Zeitpunkt beigetragen werden, sondern „longitudinal” sein, entsprechend einer regelmäßigen oder kontinuierlichen Erfassung diagnostischer und therapeutischer Informationen. Eine Aggregation derart vielfältiger Forschungsdaten, für eine große Anzahl von Individuen und zu verschiedenen Zeitpunkten erhoben, kann Ausgangspunkt für eine hocheffiziente personalisierte Gesundheitsvorsorge und eine wesentlich frühere und gezieltere Behandlung von Erkrankungen sein. In diesem Zusammenhang ist eine deutliche Reduktion der Gesundheitskosten, z.B. für Krankheitsfolgekosten oder unwirksame Therapien, zu erwarten.

4

HERAUSFORDERUNGEN

4. Herausforderungen

4.1 Datenschutz

In welchem Rahmen Daten für die Forschung genutzt und mit anderen Stellen geteilt werden dürfen, ist den Erzeugern und potenziellen Nutzern von Daten oft nicht klar. Denn die datenschutzrechtlichen Vorschriften können verschieden ausgelegt werden und fordern häufig ergebnisoffene Abwägungen für jeden einzelnen Datensatz. Die diesen Auslegungs- und Abwägungsprozessen immanenten Rechtsunsicherheiten gefährden die Ausschöpfung von Innovations- und Forschungspotenzial. Insbesondere Datennutzungsbefugnisse und das Identifikationspotenzial bestimmter Datenarten, wie der Output bildgebender Verfahren, sind datenschutzrechtlich weiterhin kontrovers diskutierte Aspekte. Länderspezifische Unterschiede in der Rechtsauslegung tun ein Übriges. Konkrete Hilfestellungen und Richtlinien, z.B. zur praktischen Anonymisierung von unstrukturierten Daten durch Aufsichtsbehörden, stehen aus. Das im Dezember 2023 verabschiedete Gesundheitsdatennutzungsgesetz sieht eine zentrale Datenzugangs- und Koordinierungsstelle vor, welche als zentrale Anlaufstelle zumindest bürokratische Hürden für Forschende abbauen soll. Diese soll das zum BfArM gehörende Forschungsdatenzentrum Gesundheit/FDZ ermächtigen können, ab etwa Mitte 2025 pseudonymisierte Gesundheitsdaten für definierte Forschungszwecke freizugeben bzw. auch Verknüpfungen mit weiteren Registerdaten zu erstellen.

Grundlage für die datenschutzrechtliche Diskussion ist die Europäische Datenschutz-Grundverordnung (DSGVO). Sie enthält eine Legaldefinition für den Begriff der „personenbezogene Daten“ in Art. 4 Nr. 4: „Personenbezogene Daten sind alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen.“ Das Grundrecht auf informationelle Selbstbestimmung liegt dem Datenschutzrecht auf nationaler Ebene zugrunde: „Das Grundrecht gewährleistet die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen.” Die Entscheidung über die Verwendung personenbezogener Daten obliegt daher stets der betroffenen Person. Dem Datenverarbeiter kommt indes die Rolle eines für die Gewährleistung der Datenschutzrechte Verantwortlichen zu. Gesetzlich werden diese Grundsätze in der DSGVO ausgestaltet. Nach Art. 6 Abs. 1 DSGVO ist eine Verarbeitung personenbezogener Daten nur dann rechtmäßig, wenn sie auf eine der dort genannten Rechtmäßigkeitsvoraussetzungen gestützt werden kann. Die Verarbeitung personenbezogener Gesundheitsdaten unterliegt nach Art. 9 Abs. 1 und 2 DSGVO noch strengeren Voraussetzungen. Jedoch enthält die DSGVO auch sogenannte Öffnungsklauseln, z.B. in Art. 9 Abs. 2 lit. j, die in Kombination mit § 27 BDSG eine Nutzung bestimmter personenbezogener Datentypen auch ohne Einwilligung abweichend von Art. 9 Abs. 1 DSGVO zu Forschungszwecken ermöglicht, falls „die Interessen des Verantwortlichen an der Verarbeitung die Interessen der betroffenen Person an einem Ausschluss der Verarbeitung

erheblich überwiegen“. Gerade diese Bedingung bereitet – sofern sie überhaupt bekannt ist – den Forschenden bei der entsprechenden Abwägung oft Kopfzerbrechen. Dies umso mehr, als in einer zunehmend datengetriebenen personalisierten Medizin davon auszugehen ist, dass immer mehr, aber immer kleinere Subgruppen von immer umfangreicheren Datenanalysen profitieren werden. Der quantitative Nutzen zum Zeitpunkt der Initiierung eines Forschungsprojektes ist daher schwer zu beziffern.

Die DSGVO-konforme primäre Handhabung von Patienten- und Untersuchungsdaten durch die datenerhebenden und -verarbeitenden Einrichtungen (Labor, Klinik etc.) ist dabei weitgehend praktikabel geregelt. Die Bedingungen für eine ggf. sogar einwilligungsfreie Forschung an medizinischen Patientendaten sind hingegen auch in den Landesdatenschutz- bzw. Landeskrankenhausgesetzen nicht immer einheitlich formuliert.

Unsicherheiten zeigen sich auch bei der Verwendung von Einwilligungsformularen bei Übereignung von Proben an Dritte. Kliniken und Praxen verwenden hier unterschiedliche Formulierungen, häufig setzen sich diese als künftige Eigentümer ein. Ein standardisiertes Einwilligungsformular[9] für einen sog. „Broad Consent“ wurde von der Medizininformatik-Initiative (MII) in Zusammenarbeit mit Universitätskliniken, Ethikkommissionen und anderen Akteuren entworfen und mit den Datenschutzbeauftragten des Bundes und der Länder abgestimmt. Begehrt die mit diesem Formular einwilligende Person die Löschung personenbezogener Daten, muss auch die zugehörige Probe vernichtet werden. Hiervon losgelöst sind Urheber- und Verwertungsrechte zu betrachten. Verlassen Biomaterialien und Daten den geschützten Raum von behandelndem Arzt und Patient, ist eine Unkenntlichmachung der Bezugsperson notwendig. Diesbezüglich sind zwei zentrale Maßnahmen zu unterscheiden: Die Pseudonymisierung (potenzieller Personenbezug bleibt chiffriert erhalten, indem Identifikatoren wie der Name durch einen Zuordnungsschlüssel ersetzt werden) und Anonymisierung (Identifikatoren der natürlichen Person werden vollständig entfernt, so dass die Person nicht mehr identifizierbar ist). Während eine Pseudonymisierung als technische Maßnahme zu qualifizieren ist, die die Identifizierung der betroffenen Person für Unbefugte erschwert, führt die Anonymisierung zur Entfernung des Personenbezugs und damit zur Unanwendbarkeit des Datenschutzrechtes. Die Daten dürfen dann frei verarbeitet und geteilt werden.

Uneinigkeit besteht in der Frage, welche Datenarten überhaupt ausreichend sicher anonymisiert werden können, insbesondere im Kontext genetischer Analysen. Laut Auskunft des Landesbeauftragten für Datenschutz und Informationsfreiheit Mecklenburg-Vorpommern vertreten die Datenschützer diesbezüglich folgende Position: „Auch wenn es Bestrebungen gibt, die Anforderungen an die Anonymisierung im Forschungsbereich aufzuweichen, halten wir daran fest, dass es sich nur dann um anonyme Daten handelt, wenn weder der Forscher noch Dritte einen Personenbezug

herstellen können. Im Einzelfall muss entschieden werden, ob Daten aus bildgebenden Verfahren überhaupt anonymisiert werden können. Biomaterialien und genetische Daten können nach unserem Verständnis, das weit überwiegend auch von den Kolleginnen und Kollegen in der Datenschutzkonferenz (DSK) geteilt wird, weder anonymisiert noch pseudonymisiert werden.“ Ein neues EuGH-Urteil legt hingegen deutlich weniger strenge Maßstäbe an (siehe Rn. 43-50 in[10]). Solange sich hier aber keine gefestigte Rechtsprechung herausgebildet hat und auch Leitlinien der DSK oder einzelner Datenschutzbeauftragter fehlen, wird diese Rechtsunsicherheit bleiben.

Gerade im Bereich genetischer Daten ist eine differenzierte Betrachtung erforderlich. So ist das theoretische Identifikationspotenzial bei vollständigen Genomdaten höher als bei lokalisierten, tumorspezifischen Mutationen. Dies gilt jedoch auch nur bei Zugriff auf nicht-anonymisierte genetische Referenzdaten inklusive deren hochkomplexer Interpretation. Somit erscheint der Aufwand an Zeit, Kosten und Arbeitskraft, der für die Identifikation einer natürlichen Person durch Dritte auf Basis der Analyse genetischer Daten nötig wäre, unverhältnismäßig groß bis gänzlich unrealistisch. Hinzu kommt, dass Daten einzelner Personen bei KI-basierter Forschung in der Gesamtheit der meist riesigen Datensätze „verschwinden”, so dass die gezielte Re-Identifikation maßgeblich erschwert ist. Insofern kann hier von faktischer Anonymisierung gesprochen werden, die gerade dadurch gekennzeichnet ist, dass eine Re-Identifizierung der personenbezogenen Daten nur durch einen unverhältnismäßig hohen Aufwand möglich wäre. Die faktische Anonymisierung als solche steht auch im Einklang mit Erwägungsgrund 26 S. 3, 4 DSGVO und stellt somit einen nach den Maßstäben der DSGVO konformen Anonymisierungsprozess dar[11]. Gerade vor dem Hintergrund, dass eine Re-Identifikation im Hinblick auf Daten bildgebender Verfahren als in erheblichem Maße unwahrscheinlich gilt, kann also angenommen werden, dass eine Anonymisierung möglich ist. Zudem liegt das Interesse der Verarbeitung gerade in der Erkennung übergeordneter Muster und nicht im Datensatz des Individuums, was sich im Fall der Anwendung der DSGVO positiv auf eine Abwägung zugunsten der wissenschaftlichen Forschungszwecke auswirken könnte (Art. 9 Abs 2 lit. j) i.V.m. § 27 I BDSG. Aktuell wird für die Umsetzung der o.g. neuen Gesetzgebungen für Genomdaten eine Opt-in- statt einer Opt-out-Regelung angedacht, die dem speziellen Charakter genetischer Daten Rechnung tragen soll, andererseits aber den Forschungszugang zu diesem wertvollsten und relevantesten aller Datentypen in der Praxis deutlich behindern wird.

Daneben gibt es Sicherheitsmaßnahmen wie eine geeignete technische Architektur, frühzeitige Verschlüsselung durch externe Treuhänder[12], Verbleiben der Daten in definierten Forschungsdatenbanken und eine rechtliche Verpflichtung aller Projektbeteiligten. Insgesamt scheint bei konsequentem Vorgehen auch im Vergleich zur klinischen Routine ein sehr hoher Sicherheitsstandard in einem Forschungsszenario abbildbar. Inwiefern das Vorliegen zuverlässig anonymisierter klinischer Daten in Plattformen mit hochwertiger Sicherheitsarchitektur ein zusätzliches Risiko gegen-

über den in Kliniken und Praxen in Klartext vorhandenen Patientenangaben darstellt, erscheint dabei sehr fraglich. Davon unberührt bleiben aber das verbriefte Recht des Patienten und der weitgehende datenschutzrechtliche Einwilligungsvorbehalt in Bezug auf die Verarbeitung von Gesundheitsdaten. Aktives Einverständnis vs. Opt-out-Regelung, Recht des Individuums vs. gesellschaftliche Verpflichtung, Datenschutz vs. Datensolidarität: Diese Begriffe sind Eckpfeiler einer teils in philosophischen Dimensionen ausgetragenen Diskussion. Beim Blick auf die zwölfseitige „Broad Consent"-Version (Erläuterung und Einverständniserklärung) der MII (s.o.) stellt sich die Frage, ob derart komplexe Formulare geeignet sind, eine „informierte Entscheidung" auf Patientenseite herbeizuführen und dabei die Bedeutung eines Ja oder Nein im Rahmen einer gesamtgesellschaftlichen Verantwortung zu reflektieren. Auch das Anwählen von Opt-In- oder Opt-Out-Verfahren in digitalen Apps (ePA etc.) durch zumeist ältere PatientInnen könnte sich als barriere-behaftet erweisen (als Alternative soll ein Widerspruchverfahren über Ombudsstellen der Krankenkassen angeboten werden). Die Rolle des Staates (bzw. mit Blick auf den EHDS: der EU-Legislative) ist in diese Diskussion mit einzubeziehen: Ist eine Vereinfachung des Einverständnisprozesses oder die grundsätzliche Ermutigung der PatientInnen zur Beisteuerung anonymisierter Daten ein Widerspruch zur Rolle als Bewahrer der Individualrechte? In vielerlei Hinsicht erscheinen systematische Aufklärungskampagnen für die breite Öffentlichkeit in Kombination mit einer Opt-out-Regelung sinnvoller als aufwändige individuelle Aufklärungsgespräche.

Inwiefern ein EU-weit einheitliches Vorgehen mittelfristig realistisch ist, bleibt abzuwarten. Mit Blick auf die europäischen Vereinbarungen zum EHDS und den im März 2024 erzielten Minimalkonsens zeichnen sich z.B. für das Opt-out-Verfahren oder den Umgang mit besonders sensiblen Daten uneinheitliche Vorgehensweisen in den verschiedenen Mitgliedsstaaten ab.

Zu beachten ist ferner, dass die Daten oft von privatwirtschaftlichen Akteuren (Laboren, Pharmafirmen) und nicht von ärztlichen Institutionen wie Kliniken erzeugt bzw. später für die Forschung genutzt werden. Die hochspezifische Expertise verlagert sich zunehmend in die Industrie und ist dort Wachstums- und Innovationstreiber. Auch für die breite Nutzbarmachung resultierender KI-gestützter Diagnostik und Therapie erscheinen verantwortungsvolle Kommerzialisierungskonzepte unumgänglich. Eine erneute Klärung oder sogar Neubewertung der Zugangsrechte an Patientendaten bzw. allgemein die Überprüfung alter rechtlicher Formulierungen erscheint in einer vielfach „disruptiv" genannten Zeit geboten.

4.2 Technische und personelle Limitationen

Für das Trainieren robuster KI-Modelle ist das Einfließen großer Datenmengen aus vielen verschiedenen Quellen essenziell[13,14]. Um das medizinische Potenzial der vorhandenen Gesundheitsdaten zu nutzen, ist deren Aggregation in Cloud-basierten Plattformen, die mit ausreichend Speicher- und Rechenkapazität ausgestattet sind, unerlässlich. Dort können Daten geteilt und von anderen Forschungsgruppen eingesehen und genutzt werden. So kann insbesondere die Generalisierbarkeit eines Modells auf diverse Patientenpopulationen verbessert werden[15]. Kommerzielle Anbieter aus den USA (z.B. AWS, Google, Microsoft) und Europa (beispielsweise Aleph Alpha, IONOS, STACKIT, Hetzner) bieten entsprechende Lösungen für Datenspeicherung und/oder Training von KI-Modellen an. Derzeit entstehen auch internationale Initiativen und institutionelle Plattformen, wie der gemeinsame EHDS, oder die in Bayern geplante „Bavarian Cloud for Health Research". Voraussetzung für das Teilen von Daten ist deren sorgfältige Anonymisierung oder Pseudonymisierung.

Das klassische, überwachte Training eines KI-Modells erfordert eine inhaltlich homogene Datenerhebung und Datenspeicherung an der Primärquelle. Im zweiten Schritt wären eine (perspektivisch ggf. kontinuierliche/automatisierte) Einspeisung autorisierter Daten in spezifische Cloudspeicher und eine standardisierte Prozessierung unter definierten Fragestellungen erforderlich[16]. Rein technisch lassen sich Daten beim sogenannten „Confidential Computing" nicht nur im Ruhe-/Speicherzustand, sondern auch während der Prozessierung zuverlässig vor Unbefugten schützen. Sogar die temporäre Prozessierung multi-institutioneller Daten in geschützten Umgebungen ohne Dateneinsicht der eigentlichen Forschungsakteure ist möglich.

Insgesamt stellt das Fehlen einheitlicher digitaler Strukturen und Instrumente bzw. das hohe Maß an Heterogenität einen kritischen Engpass dar. Die elektronische Patientenakte (ePA) könnte bei entsprechender Nutzung ein Lösungsbaustein sein. Entwickler kommerzieller Klinik-, Arzt- und Laborinformationssysteme waren bisher nicht zu einer ausreichenden Standardisierung verpflichtet. Vielmehr versuchen sie, die Marktposition und die technischen Standards ihrer zum Teil veralteten Systeme zu schützen, den Wechsel zu anderen Anbietern zu erschweren oder sich die Einbindung komplementärer Systeme über geeignete Schnittstellen teuer vergüten zu lassen. Eine leichtgängige Interoperabilität verschiedener Systeme wird so verhindert und die Nutzung von Daten für KI-Modelle deutlich erschwert. Inwiefern z.B. bei der Umsetzung des Digitalgesetzes ausreichender Druck auf die Softwarehersteller ausgeübt wird, den neuen Verpflichtungen zur Interoperabilität nachzukommen, bleibt abzuwarten. Eine sorgfältige Erhebung und Kuratierung ausreichend differenzierter und tiefgehender diagnostischer Datensätze an einem Ort ist eine weitere Grundvoraussetzung für die Schaffung von Trainingsdatenpools für klassische KI-Modelle. Technisch anspruchsvollere Lösungen wie Swarm Learning[17] oder Federated Learning[18], bei denen keine Daten,

sondern lediglich Modellparameter geteilt werden, befinden sich derzeit in der Erprobung und bieten für die Zukunft spannende Möglichkeiten zur effizienten Nutzung von Gesundheitsdaten. Durch die Anwendung großer KI-Modelle (z.B. Foundation Models), steigen jedoch auch die Anforderungen an die Rechenkapazitäten. Das Trainieren großer Sprachmodelle wie chatGPT oder klinisch orientierter Derivate erfordert Hardware und Energie, die mit Kosten in Millionenhöhe einhergehen.

Zu diskutieren ist auch die Schaffung adäquater Anreize für die Inkaufnahme eines entsprechenden Arbeitsaufwands für die Erhebung, Strukturierung und Übermittlung entsprechender Daten in Krankenhäusern und Arztpraxen, sowie für die Bereitstellung der Daten durch Patienten. Angesichts der teils sehr hohen Marktbewertung von Firmen, deren Geschäft auf der Nutzung von Gesundheitsdaten beruht, erscheinen Modelle für eine signifikante finanzielle Beteiligung sowohl der klinischen Institutionen als auch der Patienten selbst nicht unangemessen. Nicht zu unterschätzen ist der Mehraufwand für das medizinische Personal, sowohl für die Datenerhebung als auch für eine ggf. erforderliche Patientenaufklärung. Hier zeigt sich aktuell im Zusammenhang mit der ePA-Einführung immer wieder – auch berechtigter – Unmut z.B. im Lager der ambulanten Gesundheitsversorger. Die grundsätzliche Unterstützung der Ärzteschaft für die Nutzung von Gesundheitsdaten in der Forschung ist sehr hoch, gebunden jedoch an die Forderung nach einer Abfederung dieser zusätzlichen Arbeitsbelastung [22].

5

DATENTYPEN

5 Datentypen

Die in der medizinischen Diagnostik und speziell der Hämatologie und Onkologie erhobenen Datentypen unterscheiden sich hinsichtlich ihres Identifikationspotenzials, der Datenzugangsrechte, der Zustimmungspflicht und ihrer Größe. Allen Datentypen gemeinsam ist, dass ihnen im Rahmen einer KI-basierten Prozessierung in Forschung und Routinediagnostik erhebliches Potenzial beigemessen wird, um z.B. Diagnosen lebensbedrohlicher Erkrankungen schneller, sensitiver und mit höherer Sicherheit zu stellen oder das Auftreten der Erkrankung sogar vorherzusehen, exaktere Prognosen über den Verlauf der Erkrankung zu geben, oder Therapieansprechen vorherzusagen und zielgerichtete Therapien zu ermöglichen („personalized medicine").

5.1 Histopathologie und hämatologische Zytomorphologie

5.1.1 Klinische Bedeutung

In der Leukämie- und Krebsdiagnostik sind zwei Arten mikroskopischer Gewebeuntersuchungen essentiell: die histologische Untersuchung von Tumorgewebe und die zytomorphologische Analyse von Blut- und Knochenmarkzellen. Bezüglich der Datenformate, des KI-Potenzials sowie der datenspezifischen rechtlichen Aspekte bestehen große Überschneidungen zwischen diesen beiden Typen, die daher hier gemeinsam diskutiert werden.

In der Histopathologie werden krankhafte Veränderungen anhand mikro-morphologischer und immunhistochemischer Eigenschaften an Gewebsschnitten untersucht. Tumordiagnosen können erst durch die pathologische Untersuchung bestätigt und einem Subtyp zugeordnet werden. Die zytomorphologische Beurteilung von Blut- und Knochenmarkausstrichen unter dem Mikroskop ist aufgrund ihrer schnellen Durchführbarkeit meist der erste Schritt zur Diagnose oder zum Ausschluss einer Leukämie. Dafür ist jedoch eine hohe Expertise des Untersuchers erforderlich. Insbesondere ist die Zytomorphologie geeignet, zeitnah über sinnvolle weitere Untersuchungen zu entscheiden, welche oft aufwändig und kostspielig sind, aber je nach Bedarf schnellstmöglich in die Wege geleitet werden müssen.

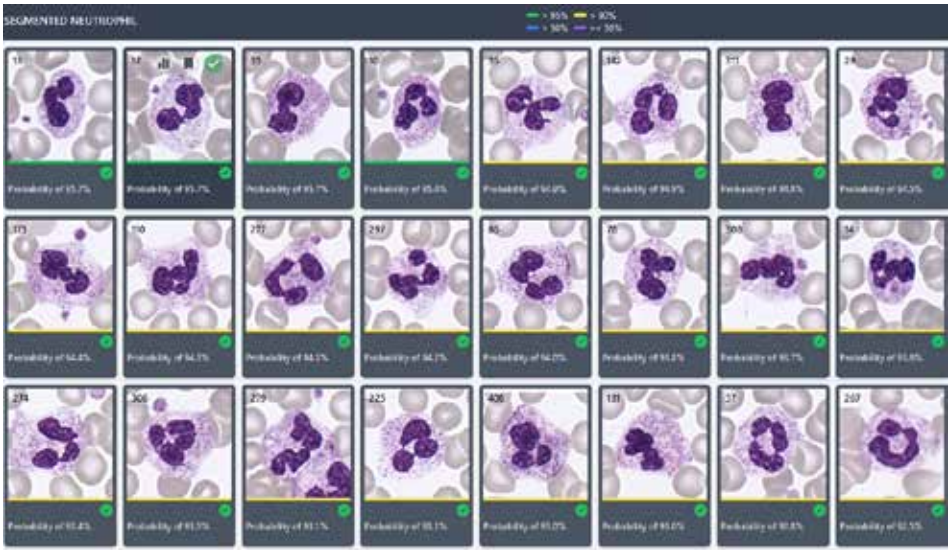
5.1.2 Daten

Histologische Schnitte und Ausstrichpräparate können in Form von Whole Slide Images (WSI) oder als Einzelaufnahmen digitalisiert werden, was im Rahmen telepathologischer Anwendungen bereits stellenweise geschieht. Sowohl Digitalisierung als auch Speicherung und Verwaltung der großen Datenmengen (pro Objektträger mind. etwa 1 Gigabyte) erfordern eine entsprechende IT-

Infrastruktur. Im Unterschied zur Zytomorphologie werden in der Histopathologie nicht nur morphologische Eigenschaften einzelner Zellen, sondern das vollständige Gewebsmuster zur Beurteilung herangezogen. Damit fließen vergleichsweise viele morphologische Parameter in die Diagnosefindung ein.

Für die klassische zytomorphologische Begutachtung werden 100-200 (kernhaltige) Zellen im Blut und 200-500 im Knochenmark klassifiziert und in die Diagnose einbezogen. Digitalisierte Datensätze umfassen daher hunderte von hochaufgelösten Einzelzellbildern (siehe Abb. 2) pro Patient. Die Größe eines solchen Datensatzes beträgt je nach Auflösung etwa 1-100 Megabyte. Einige Datensätze sind öffentlich verfügbar und entsprechend pseudonymisiert oder anonymisiert[5,23,24]. Auch hier gewinnen WSI an Bedeutung, bei denen komplette Präparate anstelle einer limitierten Anzahl von Einzelzellen hochauflösend digitalisiert werden. Hier ist der Informationsgehalt und damit die Nutzbarkeit für KI-basierte Analysen deutlich erhöht, während eine Verarbeitung aller vorhandenen Informationen – wie auch in der Histologie – durch einen Menschen undenkbar ist.
Kommerzielle Systeme zur Digitalisierung zytomorphologischer und histologischer Präparate inklusive KI-basierter Vorbefundung sind punktuell bereits verfügbar. Allerdings mangelt es noch an einem Konsens und geregelten Infrastrukturen, beispielsweise für Datenspeicherung und -austausch oder die Bildformate.

Abbildung 2: Durch KI klassifizierte weiße Blutkörperchen. Aus dem User Interface des Münchner Leukämielabors



Grafik: MLL Münchner Leukämielabor

5.1.3 KI-Potenzial

KI-Modelle können bei wichtigen histologischen Anwendungen helfen. Quantitative Bewertungen wie Proliferationsindizes in Tumoren oder Expressionsstärken von Zelloberflächenmarkern für moderne Immuntherapien sind zeitaufwändig, aber in hohem Maße automatisierbar. Auch die Feststellung des Differenzierungsgrads von Tumoren, die Untersuchung auf Lymphknotenmetastasen oder die Einteilung von Darmpolypen sind mittels KI sowohl in Geschwindigkeit als auch Genauigkeit drastisch optimierbar.

In der Zytomorphologie können KI-Algorithmen Einzelzellen auf mindestens ähnlich hohem Niveau wie erfahrene ExpertInnen klassifizieren. In Blutausstrichen konnten so verschiedenste Zelltypen unterschieden, pathologische von gesunden Zellen differenziert[25], und spezifische leukämische Subtypen[4,5] identifiziert werden. Bei der Einzelzell-Klassifikation[26] in Knochenmarkausstrichen übertreffen KI-Modelle bereits ältere, Feature-basierte Ansätze und ermöglichen auf Whole Slide-Level bereits jetzt die Detektion einer Reihe von Leukämie-Subtypen.

Unter Routine- und Studienbedingungen[27,28] zeigten sich für KI-basierte Workflows teils deutliche Vorteile in Bezug auf Sensitivität und Geschwindigkeit der Diagnosestellung. So kann ein neuronales Netz 500 Zellen in wenigen Sekunden klassifizieren, während ein sehr erfahrener menschlicher Untersucher etwa zwei Minuten für 100 Zellen benötigt. Dies ist besonders nützlich für Hochdurchsatzszenarien oder für Fälle mit nur vereinzelt auftretenden pathologischen „signature cells", die vom Menschen leicht übersehen werden können[5].

Noch unklar ist das Potenzial von KI-Modellen, die sich der Erkennung zytomorphologischer oder histologischer Muster widmen, die aufgrund ihrer Abstraktheit oder Komplexität jenseits der Wahrnehmung menschlicher UntersucherInnen liegen. Solche Modelle könnten beispielsweise die Vorhersage von Genotypen, Erkrankungsverläufen oder Therapieansprechen ermöglichen. Bei erfolgreicher Bewältigung der derzeitigen Herausforderungen wie Erklärbarkeit und Robustheit des Modells sind umfangreiche Einsatzmöglichkeiten dieser Techniken zu erwarten. Dafür ist allerdings ein Training auf Basis riesiger multi-institutioneller Bilddatensätze unabdingbar. Mit der fortschreitenden Digitalisierung von zyto-/histomorphologischen Präparaten in der Routine, vor allem für telemedizinische Anwendungen, wächst die potenzielle Datenbasis für KI-Anwendungen stetig. Allerdings ist der Bedarf bei weitem noch nicht gedeckt. Die klinische Annotation der Trainingsdaten stellt hier eine weitere große Herausforderung dar[29].

Auch aus medizin-ökonomischer Sicht wäre es interessant, eine zentrale KI-unterstützte Expertenbefundung diagnostischer Bilddatensätze auf Cloud-Plattformen einzurichten, um dem zunehmend

dramatischen Mangel an menschlichen ExpertInnen entgegenzuwirken. Auch in KI-unterstützten Workflows bleibt aber eben diese finale Befundvalidierung durch menschliche UntersucherInnen selbstverständlich zentral.

5.1.4 Datenspezifische rechtliche Aspekte

Für sich genommen erlauben anonymisierte Bilddaten aus der Histopathologie und Zytomorphologie keinen Rückschluss auf den individuellen Patienten. Mit Blick auf ihr Identifikationspotenzial sind Einzelzellbilder, Gewebsausschnitte und WSI vergleichbar einzuordnen.

5.2. Radiologie

5.2.1 Klinische Bedeutung

Die klinische Bedeutung medizinischer Bilddaten aus der Radiologie ist groß und erstreckt sich auf viele Bereiche der Patientenversorgung. Sie enthalten eine Vielzahl diagnostischer Informationen, welche aber nicht in strukturierter Textform vorliegen und zudem in hohem Maße einer subjektiven Interpretation durch menschliche UntersucherInnen („interobserver variability") unterliegen. Die hochauflösenden Bilder moderner Scanner können selbst kleinste morphologische oder funktionelle Veränderungen in Organen und Geweben sichtbar machen. Dadurch haben sie bei der Diagnose, insbesondere der Früherkennung, ebenso wie bei der Planung und Erfolgskontrolle von Therapien entscheidenden Wert.

5.2.2 Daten

DICOM (Digital Imaging and Communications in Medicine) ist das vorherrschende standardisierte Dateiformat und Protokoll für die Speicherung und Übertragung medizinischer Bildinformationen. Es wird weltweit von radiologischen Geräten wie CT-, MRT- und PET-Scannern, Ultraschallgeräten und Bildarchiven (PACS, Picture Archiving and Communication Systems) verwendet. Ein wesentlicher Aspekt des DICOM-Standards ist, dass er mehr als nur ein Bildformat ist. Jede DICOM-Datei enthält neben dem Bild selbst eine Vielzahl von Metadaten, darunter Patienteninformationen, Angaben zum medizinischen Gerät, mit dem das Bild erzeugt wurde, und spezifische Parameter des Scanvorgangs. Infolge dieser umfangreichen Metadaten eignet sich DICOM ideal für die klinische Forschung und die Anwendung künstlicher Intelligenz, da Bilddaten unter Berücksichtigung von Metadaten geordnet werden können. Trotz dieser Vorteile bringt DICOM einige Herausforderungen mit sich, insbesondere in Bezug auf das Datenmanagement und die Datensicherheit. Diese treten sowohl auf institutioneller Ebene als auch beim Austausch von Daten zwischen verschiedenen In-

stitutionen auf. Eine wichtige Voraussetzung ist die gesicherte Interoperabilität der zur Verfügung stehenden Daten. Trotz des Standards kann es zudem in der Praxis vorkommen, dass Metadatenfelder fehlerhaft ausgefüllt oder ausgelassen werden[30]. Darüber hinaus erfordert die Verwaltung großer Mengen von DICOM-Dateien, wie sie in modernen Gesundheitseinrichtungen anfallen, robuste Speicherlösungen und effiziente Datenanalysewerkzeuge.

5.2.3 KI-Potenzial

Die Auswertung dieser Bilder ist oft eine zeitaufwändige und komplexe Aufgabe, die ein hohes Maß an Fachwissen erfordert. KI-basierte Systeme haben das Potenzial, diese Prozesse zu optimieren, indem sie z.B. die Bildanalyse und -interpretation von modernen bildgebenden Verfahren wie CT, MRT und PET automatisieren. Dies könnte nicht nur dazu beitragen, die Genauigkeit von Diagnosen zu verbessern, sondern auch das medizinische Personal zu entlasten und damit die Patientenversorgung insgesamt zu verbessern. KI-Methoden können beispielsweise bei der automatischen Erkennung und Charakterisierung von Tumoren helfen, indem sie Muster und Merkmale in Bilddaten identifizieren, die für das menschliche Auge möglicherweise schwer zu erkennen sind. Sie können auch eingesetzt werden, um den Verlauf einer Krankheit oder die Reaktion auf eine Behandlung zu verfolgen, indem Veränderungen im Laufe der Zeit genau quantifiziert und dokumentiert werden[33]. Für die quantitative Erfassung von Prozessen (Größenzunahme/-abnahme, Durchblutung, Stoffwechselaktivität) sind KI-Modelle menschlichen Untersuchern hochüberlegen, insbesondere mit zunehmender Komplexität der untersuchten Sachverhalte. Für die Forschung bietet die KI das Potenzial, neue Einblicke in Entstehung, Fortschreiten und individuelle Prognose von Krankheiten zu gewinnen.

5.2.4 Datenspezifische rechtliche Aspekte

DICOM-Dateien enthalten sowohl in den Meta- als auch in den Bilddaten sensible Gesundheitsinformationen und grundsätzlich potenziell identifizierende Informationen (z.B. zur Gesichtsrekonstruktion). Pseudonymisierungs- und Anonymisierungsverfahren und ein sicherer Datentransfer sind daher unerlässlich. Gleichzeitig muss jedoch sichergestellt werden, dass die entfernten oder geänderten Daten nicht für die klinische oder wissenschaftliche Analyse der Bilder erforderlich sind.

5.2.5 Ausblick

Insbesondere in Bezug auf komplexe oder seltene Erkrankungen ist die Verfügbarkeit von Trainingskohorten mit mehreren tausend Datensätzen innerhalb einer Institution nicht gegeben. Mögliche Lösungsansätze stellen methodische (z.B. privatsphärenwahrendes verteiltes Lernen[32])

und strukturelle (z.B. Etablierung von Forschungsverbünden) Entwicklungen dar. Das Radiological Cooperative Network, RACOON[34], stellt eine deutschlandweite Forschungs- und Versorgungsinfrastruktur dar, die unter Einhaltung ethischer und datenschutzrechtlicher Grundsätze eine Nutzung von klinischen Bild- und Metadaten deutscher Universitätskliniken für die sichere Entwicklung von KI-Systemen ermöglicht.

5.3. Genetik

5.3.1 Klinische Bedeutung

Genetische/genomische Daten stellen den vermutlich wertvollsten Datentyp in der Gesundheitsforschung dar. Für fast alle bösartigen Erkrankungen sind genomische Daten ein essenzieller Bestandteil der Diagnostik, Therapieplanung und Prognoseabschätzung. Es wird zunehmend versucht, die Diagnose und Prognose ausschließlich anhand genetischer Merkmale des Tumorgewebes zu definieren. Auch für den Einsatz zielgerichteter Medikamente in der „personalized medicine" muss eine genetische Diagnostik vorangestellt werden. Stehen keine etablierten Therapien mehr zur Verfügung, wird im „Molekularen Tumorboard" nach tumorindividuellen genetischen Charakteristika gesucht, die lediglich die Tumorzellen selbst betreffen und z.B. die Aggressivität der Erkrankung oder die Prognose und das Therapieansprechen bestimmen. So kann eine rationale und personalisierte Therapie ermöglicht werden. Genetische Daten werden auch zur Verlaufskontrolle von Leukämien und Tumorerkrankungen herangezogen, um die messbare Resterkrankung zu quantifizieren. Während die bisher erwähnten genetischen Charakteristika erworbene/somatische Varianten betreffen, sind angeborene oder Keimbahnvarianten z.B. relevant für die Erkennung eines familiären oder individuellen Krebsrisikos oder für die Verstoffwechslung von Medikamenten. Das ursprüngliche, „vererbte" Genom eines Menschen beinhaltet letztlich den hochkomplexen Code für all seine individuellen biologischen Merkmale. Aktuell wird intensiv daran gearbeitet, genetische Charakteristika für die Früherkennung von Krebs einzusetzen oder sogar Tumorerkrankungen vorzubeugen. Auch für die meisten anderen Erkrankungen sind direkt oder indirekt kausale Zusammenhänge mit individuellen Merkmalen des Erbguts anzunehmen. Selbst die Verträglichkeit von Medikamenten, Noxen oder Nahrungsmitteln wird oft genetisch determiniert.

5.3.2 Daten

Die Daten bestehen meistens aus der Nukleotidsequenz der DNA, seltener der RNA oder anderen Eigenschaften der DNA. Die Nukleinsäuren werden meist aus den Tumorzellen oder als zellfreie DNA aus Blutplasma isoliert. Teilweise wird zur Kontrolle auch DNA aus gesundem Gewebe verwendet. Der Informationsgehalt eines menschlichen Genoms entspricht etwa 750 MB. Genetische

Informationen können mit unterschiedlichen Methoden erhoben werden, darunter Chromosomenanalyse, PCR (Polymerase-Kettenreaktion) und Next-Generation-Sequencing (NGS). Die meisten Daten werden heutzutage mittels NGS generiert, von definierten Genabschnitten bis hin zum kompletten Genom. Für das Speichern von Rohdaten und Qualitätsmerkmalen und den Abgleich individueller Gensequenzen mit Referenzen aus großen Vergleichspopulationen in internationalen Datenbanken existieren hochgradig standardisierte Dateiformate und Softwaretools. Genomdaten werden in der Regel bei der Sequenzierung vielfach gelesen, um valide und ausreichend sensitive Ergebnisse zu gewährleisten. Hieraus resultieren Datensätze von bis zu 500 GB pro Individuum. In der Forschung finden Datensätze Anwendung, die Genomdaten hunderter oder tausender Patienten enthalten, einschließlich zusätzlicher Informationen zu Diagnose, Therapieansprechen, oder auch komplexer mikroskopischer oder radiologischer Bilddaten.

5.3.3 KI-Potenzial

Bei genomischen Daten in der Medizin handelt es sich um sehr große, sequenziell strukturierte Datenmengen, die sich für die KI-basierte Analyse z.B. durch Transformermodelle (wie bei den großen Sprachmodellen) anbieten[35].

Bisher wird nur ein kleiner Teil der vorhandenen Daten für die klinische Bewertung genutzt und viele Zusammenhänge sind noch völlig unklar. Daher ist das Potenzial von KI bei dieser Datenform besonders groß. Es ist sehr wahrscheinlich, dass KI die Diagnostik anhand genetischer Daten deutlich verbessert und beschleunigt, Krankheitsverläufe vorhersagen kann, die Toxizität von Medikamenten individuell einschätzen kann, eine wichtige Rolle bei der Prävention von Erbkrankheiten spielen wird und für die Entwicklung neuer Medikamente essentiell ist[36]. Auf Ebene des einzelnen Patienten bietet die KI die Möglichkeit, personalisierte Behandlungspläne zu entwickeln, die auf den individuellen biologischen Merkmalen des Patienten und der Art des Tumors basieren. Dies kann dazu beitragen, die Wirksamkeit der Behandlung zu erhöhen und Nebenwirkungen zu reduzieren[37,38].

Ein hohes Potenzial ist bereits in der Biomarker-Forschung erkennbar. Es wird derzeit auch intensiv daran gearbeitet, eine Datenbank aufzubauen, in der digitale Zwillinge (Digital Twins) hinterlegt sind, anhand derer wirksame Therapien vorhergesagt werden können. Aus medizinökonomischer Sicht zeichnet sich ein immenses Potenzial für eine Rationalisierung medizinischer Diagnostik und Therapie ab, inklusive der Vermeidung eines erheblichen Teils von Krankheitsfolgekosten. Dieses Potenzial lässt sich aber nur durch eine großflächige Nutzung genomischer Daten in Kombination mit den zugehörigen diagnostischen und therapeutischen Patientendaten vollständig realisieren und ist durch die aktuelle rechtliche Handhabung genetischer Daten deutlich limitiert.

5.3.4 Datenspezifische rechtliche Aspekte

Genetische Daten gelten im Sinne der Datenschutzgrundverordnung (DSGVO) als sensible personenbezogene Daten. Im Rahmen der medizinischen Versorgung und Forschung ist die Verwendung genetischer Daten zu Diagnose-, Behandlungs- und Forschungszwecken zulässig, sofern die Vertraulichkeit und Integrität der Daten gesichert sind und die Rechte und Interessen der betroffenen Personen (inkl. Einwilligung) gewahrt werden. Genomische Daten sind – je nach Umfang – dazu geeignet, ein Individuum eindeutig zu identifizieren, sofern auf genetische Referenzdaten desselben Individuums (oder von Verwandten) zugegriffen werden kann oder diese aus vorhandenem Biomaterial zum Abgleich erneut erhoben werden können. Ohne diese Referenzdaten ist ein genomischer Datensatz – auch dies ist zu beachten – nicht für die Identifikation eines Individuums geeignet, es sei denn, es besteht Zugriff auf genetische Daten enger Verwandter, die entsprechende Rückschlüsse erlauben. Trotzdem ist in den neuen Gesetzgebungen die Sonderstellung genetischer Daten unterstrichen, die weiterhin einer aktiven (Opt-in) statt einer passiven (Opt-out) Einwilligung für die Forschungsverwendung bedürfen.

6

DISKUSSION

6. Diskussion

6.1 Charakteristika medizinischer Datentypen

Wie aus den oben aufgeführten Anwendungsbeispielen ersichtlich wird, existieren sehr unterschiedliche medizinische Datentypen, die allesamt dazu geeignet sind, die KI-basierte medizinische Forschung zu beschleunigen. Dabei können sie eine Vielzahl völlig neuer, potenziell disruptiver Erkenntnisse für Diagnostik und Therapie im Rahmen einer personalisierten Medizin liefern. Langfristig ist das plausible Ziel, allen PatientInnen die wirksamste und verträglichste Therapie auf Basis ihrer individuellen Erkrankungstypen und -stadien, ihrer genetisch determinierten biologischen Eigenschaften und der spezifischen genetischen Eigenschaften ihrer Krebserkrankungen innerhalb kürzester Zeit anbieten zu können. Aufgrund vieler noch unverstandener Zusammenhänge insbesondere im Bereich des Genoms, ist das volle Ausmaß von möglichen langfristigen Innovationen für die Patientenversorgung (nicht nur im Bereich der Hämatologie/Onkologie) zum aktuellen Zeitpunkt nicht wirklich abschätzbar. Das Outcome einer solchen Daten- und KI-getriebenen Forschung beruht ganz wesentlich auf der Qualität und damit der Auffindbarkeit, dem Umfang, der Tiefe bzw. Differenziertheit und der Annotation der genutzten Daten. Zur Gewährleistung entsprechender Datenqualität ist zunächst sicher ein signifikanter Mehraufwand an finanziellen, personellen und technischen Ressourcen einzukalkulieren. In der Folge kann jedoch ein dauerhaft entlastender Effekt auf die Budgets der Gesundheitsökonomie erwartet werden.

6.2 Juristischer Kontext

Neben technischen Limitationen, die zu einem Großteil temporär erscheinen, verhinderten bislang insbesondere datenschutzrechtliche Aspekte die volle Ausschöpfung des Datenpotenzials. Viele der rechtlichen Barrieren dienen dabei einem wichtigen und unverhandelbaren höheren Ziel: dem Schutz vor Missbrauch intimer persönlicher Daten. Die Zukunft der medizinischen Forschung und Entwicklung, sowohl institutionell als auch privatwirtschaftlich, wird entscheidend davon abhängen, ob es gelingt, eine praktikable und konsensfähige Abwägung zwischen der Gesetzeslage und ihrer Auslegung zu finden. Diese Abwägung sollte realistische Gefahren pragmatisch und objektiv auf ein Minimum beschränken, andererseits aber das große Potenzial der KI-basierten Forschung an Gesundheitsdaten für Gesellschaft und Allgemeinwohl nicht unnötig einschränken. Speziell in Deutschland wird die Diskussion teils sehr emotional geführt und ist zugleich von Unsicherheit geprägt. Eine länderübergreifende Vereinheitlichung des Umgangs mit Gesundheitsdaten könnte angesichts einer föderal geprägten Datenschutzkultur mehr Klarheit schaffen. Die Verhältnismäßigkeit der primären und sekundären Handhabung von Gesundheitsdaten im Vergleich zu vielen anderen persönlichen Daten (Social Media, Online-Shopping etc.) scheint aber auch unabhängig

davon an vielen Stellen verzerrt zu sein. Staatlich geschützte und garantierte Datenplattformen erscheinen geeignet, hier Vertrauen zu schaffen. Die seit langem geforderte Etablierung solcher Systeme kommt jedoch noch deutlich zu langsam voran.

Um eine Balance zwischen der Nutzung der Daten und der Wahrung der Datenrechte zu finden, sind mehrere Strategien denkbar:

- Unterschiedliche Sicherheitslevel für Datensätze, je nachdem, ob z.B. genetische Daten mit Patienten- oder Krankheitscharakteristika oder klinischen Verläufen assoziiert sind oder nicht.
- Daten-Treuhandmodelle für den Fall, dass zentrale Datenauswertungen erforderlich sind.
- Föderiertes Lernen, bei dem nur die aggregierten Analysen geteilt werden, aber nicht die Primärdaten.
- Homomorphe Verschlüsselung, bei der die Daten analysiert werden, ohne sie für den Nutzer zu entschlüsseln.

Die oben genannten nationalen Gesetzespakete werden ebenso wie die Strukturen des EHDS – zumindest in der Theorie – den Zugang zu Gesundheitsdaten künftig erleichtern. Die bereits auf verschiedenen Ebenen diskutierte Idee eines Gesundheitsdaten-Solidarpakts, der die Nutzung pseudonymisierter bzw. anonymisierter Patientendaten erlaubt, sofern dem nicht explizit patientenseitig widersprochen wird („Opt-out"-Regelung), soll mittels der neuen Gesetzgebung schrittweise umgesetzt werden und ist eine aus unserer Sicht absolut unterstützenswerte Lösung. Jedoch sind genetische Daten von dieser Regelung ausgenommen und bedürfen – so der aktuelle Plan – für die Forschungsnutzung weiterhin einer expliziten Einwilligung („Opt-in"-Regelung). Begrüßenswerte Initiativen wie die Nationale Strategie für Genommedizin/genomDE und das darauf beruhende „Modellvorhaben Genomsequenzierung" sind ein wichtiger Schritt in die richtige Richtung und zielen auf eine Verwendung von Whole Genome-Daten nicht nur in der primären Diagnostik, sondern auch im Forschungskontext ab. Ein barrierefreier Zugang zu genetischen Daten ist allerdings in Deutschland und Europa nicht absehbar, da das Recht auf informationelle Selbstbestimmung der betroffenen Personen hier weiterhin Priorität genießt[39]. Dies erschwert eine Nutzung des vollen Potenzials der Gesundheitsdaten erheblich. In diesem Zusammenhang sollten fortwährend und in regelmäßigen Intervallen die Rechte und Interessen des Individuums gegenüber dem Allgemeinwohl bzw. den Interessen der Gesellschaft abgewogen werden und das kontinuierlich und drastisch wachsende Potenzial von Forschung und Technologie möglichen Nachteilen und Missbrauchsszenarien gegenübergestellt werden. Dabei sind unbedingt der anhaltend rasante technische Fortschritt und die sich daraus immer wieder neu ergebenden großen Möglichkeiten auf aktuellem Stand in die Diskussion der Entscheidungsträger einzubringen. Der entsprechende Dialog ist

zwingend auf europäischer Ebene zu führen und müsste die datenschutzrechtlichen Vorgaben der Einwilligungsverfahren dezidiert für den Gesundheitsbereich betrachten.

Bezüglich der Opt-in-Regelung für genetische Gesundheitsdaten gilt zudem: auf Patientenseite sollte der Prozess einfach, verständlich und möglichst universal handhabbar sein, ohne die Selbstbestimmung der PatientInnen signifikant zu beeinträchtigen, aber auch ohne die Bedeutung der Forschung an genetischen Daten zu nivellieren. Erfreulicherweise werden nach den jüngsten Gesetzesbeschlüssen künftig keine ungerechtfertigten Differenzierungen zwischen institutionellen und privaten Forschungsakteuren mehr vorgenommen, sondern ausschließlich Grund und Ziel des Forschungsvorhabens betrachtet. Dennoch sind Unternehmen hier klar in der Pflicht, transparent und regelmäßig aufzuzeigen, dass wirtschaftlicher Profit nicht über das Patientenwohl gestellt wird.

6.3 Ökonomisches Potenzial

Nicht unerwähnt bleiben soll das grundsätzliche ökonomische Potenzial der Gesundheitsdatenforschung, welches auch bei einem ausreichend verantwortungsvoll gehandhabten Datenschutz hochsignifikant sein dürfte. Es handelt sich um ein Feld, welches mit Healthcare und AI/Data Science zwei der international zukunftsträchtigsten und disruptivsten Branchen vereint. Der Bund der deutschen Industrie BDI erwartet durch die allgemeine Digitalisierung des Gesundheitswesens für 2030 ein Wertschöpfungspotenzial von bis zu 140 Mrd. Euro allein in Deutschland. KI-gestützte Anwendungen erzeugen heute schon enorme finanzielle Werte. So wurden im Jahr 2022 drei Milliarden US-Dollar in Start-ups investiert, die KI auf medizinische Fragen anwenden[40].

Pharmaforschung ist ohne „Omics"-Daten (Genomics, Proteomics etc.) heute kaum noch vorstellbar; wobei die Grenzen zwischen biologischer und IT-Forschung verschwimmen. Zukünftig dürfte ein großer Teil des weltweiten Pharmaumsatzes von aktuell fast 1,5 Billionen US-Dollar mit Medikamenten erzielt werden, die auf Basis der Analyse von Patientendaten entwickelt wurden. Die Frage liegt nahe, wem dieser enorme Wert „gehört". Neben rechtlichen Fragen berührt das auch verteilungspolitische Themen[41]. Zur Wahrheit gehört jedoch auch, dass das immense Potenzial der Gesundheitsforschung einen Schatz darstellt, der ohne Hilfe privatwirtschaftlicher Akteure mit ihren effizient zugeschnittenen Strukturen und ihrer beachtlichen Monetarisierung nicht zu heben ist. Grundsätzlich ist kein Widerspruch darin zu sehen, mit einer am Gemeinwohl ausgerichteten Gesundheitsforschung Profite zu erwirtschaften (und damit auch Arbeitsplätze zu schaffen). Ggf. wäre sogar über ökonomische Beteiligungsmodelle für primär in die Datenverarbeitung involvierten Personen (PatientInnen, KlinikerInnen) nachzudenken, insbesondere bei Verwendung in der kommerziellen Forschung.

Weltweite Entwicklungen im Bereich der Informationstechnologie und KI-Forschung – auch im Zusammenhang mit Gesundheitsdaten – entziehen sich jedoch zunehmend einer gesellschaftlichen oder staatlichen Kontrolle. Privatwirtschaftliche und teils auch institutionelle Akteure verfolgen Ziele, die nicht einem allgemeinen ethischen oder medizinisch-wissenschaftlichen Konsens unterliegen. Kontrollorgane sind oft nur in der Lage, auf neue Entwicklungen zu reagieren, können diese aber kaum proaktiv steuern. Signifikante technische Meilensteine lösen oft aktionistische Maßnahmen aus, von Hype bis Verdammung, wie beispielsweise im Fall von ChatGPT.

Dabei entwickelt sich die Technik auf einer infrastrukturellen Basis, die dafür vielfach ungeeignet und unsicher ist und ungewollt Monopole befördert. Mit Blick auf die Geschwindigkeit der technischen Weiterentwicklungen ist anzunehmen, dass diese Problematik sich weiter zuspitzen wird. Auch unter diesem Aspekt erscheint ein breiterer, verantwortungsbewusst regulierter Zugang zur Forschung mit Gesundheitsdaten für seriöse Akteure womöglich sinnvoller als ein harter Restriktionskurs, der eher die Monopolisierung zugunsten zweifelhafter Akteure fördert. Hier wird es auf die praktische Auslegung der neuen Gesetzespakete (s.o.) in Deutschland sowie des EU AI Act und der Bestimmungen zum EHDS ankommen.

7

SCHLUSSFOLGERUNGEN

7. Schlussfolgerungen

Auf EU- und Bundesebene wurden zuletzt wichtige Gesetzgebungsvorhaben zum Thema deutlich vorangetrieben und zielen auf einen sicheren und klarer geregelten Datenzugang auch im Forschungsbereich ab. Zudem finden sich in Deutschland bereits seit Jahren etliche namhafte Initiativen, welche in unterschiedlicher Nuancierung eine flächendeckende Forschung an Gesundheitsdaten vorantreiben wollen und einen pragmatischeren und nutzenorientierteren Umgang zum Wohle aller PatientInnen anmahnen und anstreben.

So zeigt sich letztlich, dass die Nutzung von Gesundheitsdaten aus Sicht der im Forschungskontext damit befassten Akteure schon seit geraumer Zeit einen erheblichen Stellenwert besitzt. Insbesondere in der Hämatologie/Onkologie mündet dies in etlichen Initiativen hochrangiger Fachgesellschaften und Interessensvertretungen, mit denen wir uns ausdrücklich solidarisieren. Viele dieser Akteure sind bemüht, auch ohne Unterstützung von staatlichen Institutionen die vorhandenen technischen und datenschutzrechtlichen Hindernisse anzugehen. Eine gelegentlich bemängelte Frontenbildung zwischen vermeintlich fortschrittsfeindlichen Datenschützern und allzu bedenkenlosen Forschenden ist vielfach nicht wahrzunehmen. Vielmehr finden sich disziplinübergreifende Kooperationsprojekte, in denen deutlich wird, dass etliche Akteure auf beiden Seiten letztlich das gleiche Ziel verfolgen, sich aber aus unterschiedlicher Perspektive annähern. Dieser kollaborative Ansatz ist äußerst begrüßenswert. Mit Blick auf die unlängst verabschiedeten Entwürfe für das neue Gesundheitsdatennutzungsgesetz und das neue Digitalgesetz, die erwarteten Verabschiedungen des Medizinforschungsgesetzes und des Digitalagenturgesetzes sowie des AI Act der EU und des EHDS ist eine sehr frühzeitige Einbindung aller Stakeholder in konstruktive und im Geiste guter Zusammenarbeit geführte Gespräche über die praktische Umsetzung dringend angezeigt.

Die neuen Gesetze bergen zu viele potenzielle Fallstricke, als dass ein Sich-Ausruhen auf dem einmaligen legislativen Akt für Forschende, DatenschützerInnen oder Ministeriale erlaubt sein dürfte. Ferner besteht die Gefahr, dass Deutschland im europäischen Vergleich abermals zurückfällt und die Auslegung des Gesundheitsdatennutzungsgesetz der europaweit praktizierten Nutzung des EHDS hinterherhinken wird. Zwar ist die Planung einer Opt-out-Regelung für die Weitergabe von Gesundheitsdaten zu Forschungszwecken positiv zu bewerten. Auch die Idee einer zentralen Koordinations- und Zugangsstelle (als Weiterentwicklung des sog. Forschungsdatenzentrums/FZA des Bundesinstituts für Arzneimittel und Medizinprodukte/BfArM), welche aus elektronischer Patientenakte, GKV-Daten u.a. gespeist wird, prinzipiell zu begrüßen. Allerdings scheint das Ausmaß des geplanten Regulierungsaufwands bei absehbar fehlenden Personalressourcen äußerst kritisch und lässt eine effiziente administrative Bearbeitung bei erwartbar hoher Frequentierung mehr als fraglich erscheinen[42]. Einmal mehr droht hier die Gefahr einer (typisch deutschen?) Überbürokrati-

sierung. Gleichzeitig sind die für Forschende wertvollsten diagnostischen Daten, die Genomdaten, vom Opt-out-Prozess ausgeschlossen, während die Umsetzung eines Opt-in-Prozesses über eine Handy-App für zumeist ältere PatientInnen ohne sehr ausführliche Anleitung und Aufklärung problematisch sein könnte. Der geplante Zugriff der Krankenkassen auf die Daten der Versicherten zum Zwecke eines „individualisierten Angebots" für Patientinnen und Patienten hat aus Sicht manches Datenschützers einen zweifelhaften Beigeschmack.

Im Umsetzungsprozess ist es grundlegend, kontinuierlich zu hinterfragen: Wovor genau will und muss man die Patienten und ihre Daten schützen? Auf der einen Seite stehen forschende Institutionen und Firmen, die nicht an den Daten identifizierbarer Individuen interessiert sind, sondern an übergeordneten Mustern, die sich in großen Patientendatensätzen abzeichnen und die Ermittlung neuer Erkrankungssubtypen und Therapieansätze ermöglichen. Auf der anderen Seite finden sich zweifelhafte Akteure, die an Daten definierter Individuen interessiert sind und daraus Maßnahmen ableiten können. Erstere so gut wie möglich zu fördern und letztere möglichst zuverlässig fernzuhalten, mit allen technischen Mitteln, muss das gemeinsame Ziel sein. Allen verantwortlichen Akteuren muss jederzeit bewusst sein, dass ein mangelnder oder verzögerter Zugang forschender Institutionen zu Gesundheitsdaten einer weiteren Verbesserung der Patientenversorgung entgegensteht bzw. Patientenleben gefährden kann. Daher sollten entsprechende Einschränkungen sehr kritisch auf Angemessenheit und Plausibilität geprüft werden, nicht zuletzt im Sinne der vielen bereits an Krebs Erkrankten und in Zukunft Erkrankenden, deren Behandlung und Versorgung sich nicht zuletzt unser Arbeitskreis und die gesamte Deutsche Gesellschaft für Hämatologie und Onkologie verschrieben haben.

Daher setzen sich die Autoren und Unterstützer dieses Positionspapiers konkret ein für

- die Schaffung geeigneter Hardware- und Software-Infrastrukturen einschließlich potenter Cloud-Server, einheitlicher Schnittstellen und verpflichtender Standards für medizinische IT-Systeme (Klinik-, Praxis- und Laborinformationssysteme/-verwaltungssysteme), inklusive einer Überwachung der praktischen Umsetzung gemäß Digitalgesetz (z.B. durch die gematik),
- eine konsequente, quellenübergreifend standardisierte, sichere und möglichst automatisierte Erfassung und Übermittlung des juristisch vertretbaren Maximums an Gesundheitsdaten (in die elektronische Patientenakte und ggf. – pseudonymisiert – in das FDZ oder den EHDS), mit Start zum nächstmöglichen Zeitpunkt,
- die Schaffung adäquater personeller Voraussetzungen und (auch monetärer) Anreize für eine effektive Datenerhebung bzw. -weitergabe insbesondere in den Institutionen der medizinischen Primärversorgung sowie auf Patientenebene,

- eine umfassende wie einfach verständliche Aufklärung der Bevölkerung über die Bedeutung der Gesundheitsdatenforschung und die technische Datensicherung zur Schaffung breiter Akzeptanz für die Einwilligungsverfahren,
- die Nutzbarmachungung von Gesundheitsdaten für forschende Institutionen und Unternehmen in technisch nach neuesten Standards gesicherten und staatlich garantierten Verarbeitungsumgebungen und
- die Schaffung länder- und institutionsübergreifend vereinheitlichter, verständlicher Leitlinien für eine praktikable Auslegung der relevanten datenschutzrechtlichen Vorschriften, samt ausreichender administrativer Ressourcen für eine zügige Bearbeitung von Anträgen auf Forschungsdatenzugriff.

Literatur

1. Walter-Siegenthaler-Gesellschaft. Digitalisierung der Medizin für das Patienten- und Gemeinwohl. (2023). https://siegenthaler-gesellschaft.de/wp-content/uploads/2023/09/Digitalisierung-der-Medizin-fuer-das-Patienten-und-Gemeinwohl_09.08.2023.pdf
2. Walter, W. et al. Artificial intelligence in hematological diagnostics: Game changer or gadget? Blood Rev. 58, 101019 (2023). DOI: 10.1016/j.blre.2022.101019
3. Luchini, C., Pea, A. & Scarpa, A. Artificial intelligence in oncology: current applications and future perspectives. Br. J. Cancer 126, 4–9 (2022). DOI: 10.1038/s41416-021-01633-1
4. Sidhom, J.-W. et al. Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. NPJ Precis Oncol 5, 38 (2021). DOI: 10.1038/s41698-021-00179-y.
5. Hehr, M. et al. Explainable AI identifies diagnostic cells of genetic AML subtypes. PLOS Digit Health 2, e0000187 (2023). DOI: 10.1371/journal.pdig.0000187
6. Eckardt, J.-N. et al. Deep learning identifies Acute Promyelocytic Leukemia in bone marrow smears. BMC Cancer 22, 201 (2022). DOI: 10.1186/s12885-022-09307-8
7. Binder, A. et al. Morphological and molecular breast cancer profiling through explainable machine learning. Nature Machine Intelligence (2021) DOI:10.1038/s42256-021-00303-4.
8. Bjornevik, K. et al. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science 375, 296–301 (2022). DOI: 10.1126/science.abj82229.
9. Mustertext zur Patienteneinwilligung. https://www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung.
10. CURIA - Documents. https://curia.europa.eu/juris/document/document.jsf?text=&docid=272910&pageIndex=0&doclang=DE&mode=lst&dir=&occ=first&part=1&cid=1107245
11. Mühlenbeck, R. L. Anonyme und pseudonyme Daten. (Nomos, 2023). https://www.beckshop.de/muehlenbeck-anonyme-pseudonyme-daten/product/35449874
12. Specht-Riemenschneider, L. et al. Die Datentreuhand: ein Beitrag zur Modellbildung und rechtlichen Strukturierung zwecks Identifizierung der Regulierungserfordernisse für Datentreuhandmodelle. (CH Beck, 2021).
13. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. (2019) DOI:10.1038/s41591-019-0508-1.
14. Wagner, S. J. et al. Make deep learning algorithms in computational pathology more reproducible and reusable. Nat. Med. 1–3 (2022). DOI: 10.1038/s41591-022-01905-0
15. Willemink, M. J. et al. Preparing Medical Imaging Data for Machine Learning. Radiology 295, 4–15 (2020). DOI: 10.1148/radiol.2020192224

16. Ghassemi, M. et al. A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc 2020, 191–200 (2020). PMID: 32477638
17. Warnat-Herresthal, S. et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature 594, 265–270 (2021). DOI: 10.1038/s41586-021-03583-3
18. Raab, R. et al. Federated electronic health records for the European Health Data Space. The Lancet Digital Health 0, (2023). DOI: 10.1016/S2589-7500(23)00156-5
19. Truhn, D. et al. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). J. Pathol. (2023) doi:10.1002/path.6232. DOI: 10.1002/path.6232
20. Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [cs.CL] (2023). DOI: 10.48550/arXiv.2307.09288
21. Adams, L. C. et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology 307, e230725 (2023). DOI: 10.1148/radiol.230725
22. Köngeter, A., Schickhardt, C., Jungkunz, M., Mehlis, K. & Winkler, E. C. Physicians' attitudes towards secondary use of clinical data for biomedical research purposes in Germany. Results of a quantitative survey. bioRxiv (2022) DOI: 10.1371/journal.pone.0274032.
23. Matek, C., Schwarz, S., Marr, C. & Spiekermann, K. A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls (AML-Cytomorphology_LMU). The Cancer Imaging Archive (TCIA) (2019) DOI:10.7937/tcia.2019.36f509ld.
24. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T. & Marr, C. An expert-annotated dataset of bone marrow cytology in hematologic malignancies (bone-marrow-cytomorphology_MLL_Helmholtz_Fraunhofer). The Cancer Imaging Archive (TCIA) (2021) DOI:10.7937/TCIA.AXH3-T579.
25. Matek, C., Schwarz, S., Spiekermann, K. & Marr, C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat Mach Intell 1, 538–544 (2019). DOI: 10.1038/s42256-019-0101-9
26. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T. & Marr, C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. Blood 138, 1917–1927 (2021). DOI: 10.1182/blood.2020010568
27. Pohlkamp, C. et al. A Fully Automated Digital Workflow for Assessment of Bone Marrow Cytomorphology Based on Single Cell Detection and Classification with AI. Blood 140, 10725–10726 (2022). DOI: 10.1182/blood-2022-168780
28. Haferlach, T., Nadarajah, N., Haferlach, C., Kern, W. & Pohlkamp, C. Machine Learning Algorithm Correctly Identifies 95% of Cells in Differential Count of Blood Smears: A Prospective Study on ›29,000 Cases and ›17 Million Single Cells. Blood 140, 1909–1910 (2022). DOI: 10.1182/blood-2022-165863

29. Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med 4, 71 (2021). DOI: 10.1038/s41746-021-00427-2
30. Kohli, M. D., Summers, R. M. & Geis, J. R. Medical Image Data and Datasets in the Era of Machine Learning-Whitepaper from the 2016 C-MIMI Meeting Dataset Session. J. Digit. Imaging 30, 392–399 (2017). DOI: 10.1007/s10278-017-9976-3
31. Kelly, B. S. et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur. Radiol. 32, 7998–8007 (2022). DOI: 10.1007/s00330-022-08784-6
32. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence 3, 473–484 (2021). DOI: 10.1038/s42256-021-00337-8
33. Bera, K., Braman, N., Gupta, A., Velcheti, V. & Madabhushi, A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nat. Rev. Clin. Oncol. 19, 132–146 (2022). DOI: 10.1038/s41571-021-00560-7
34. Heyder, R. et al. [The German Network of University Medicine: technical and organizational approaches for research data platforms]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 66, 114–125 (2023). DOI: 10.1007/s00103-022-03649-1
35. Choi, S. R. & Lee, M. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology 12, (2023). DOI: 10.3390/biology12071033
36. Radakovich, N. et al. A geno-clinical decision model for the diagnosis of myelodysplastic syndromes. Blood Adv 5, 4361–4369 (2021). DOI: 10.1182/bloodadvances.2021004755
37. Chua, I. S. et al. Artificial intelligence in oncology: Path to implementation. Cancer Med. 10, 4138–4149 (2021). DOI: 10.1002/cam4.3935
38. Kann, B. H., Hosny, A. & Aerts, H. J. W. L. Artificial intelligence for clinical oncology. Cancer Cell 39, 916–927 (2021). DOI: 10.1016/j.ccell.2021.04.002
39. Gowda, V., Kwaramba, T., Hanemann, C., Garcia, J. A. & Barata, P. C. Artificial Intelligence in Cancer Care: Legal and Regulatory Dimensions. Oncologist 26, 807–810 (2021). DOI: 10.1002/onco.13862
40. Goldman, S. 6 AI companies disrupting healthcare in 2022. VentureBeat https://venturebeat.com/ai/6-ai-companies-disrupting-healthcare-in-2022/
41. Thielscher, C. & Kappler, K. DIGITALIZATION AND ORGANIZATION OF CARE: THE CASE OF ONCOLOGY. Eur. J. Polit. Econ. 22, 127–139 (2023). https://www.researchgate.net/publication/376650829_DIGITALIZATION_AND_ORGANIZATION_OF_CARE_THE_CASE_OF_ONCOLOGY
42. Wissenschaftspolitische Empfehlungen und Standpunkte. Fraunhofer-Gesellschaft https://www.fraunhofer.de/de/ueber-fraunhofer/wissenschaftspolitik/wissenschafts-politische-empfehlungen-und-standpunkte.html (2023).

ISBN 978-3-9821204-5-4