Fachgutachten
Machbarkeit einer deutschen MIMIC

Erstellt für das Bundesministerium für Gesundheit

Erarbeitet und vorgelegt von der Gutachtergruppe (in alphabetischer Reihenfolge):

Prof. Dr. Benedikt Buchner, Augsburg
Prof. Dr. Fabian Prasser, Berlin
PD Dr. Sven Zenker, Bonn

Version 1.1
Berlin, 12.02.2024

Inhaltsverzeichnis

Zusammenfassung

Dieses Gutachten befasst sich mit der Machbarkeit und dem Wert eines deutschen Pendants zur MIMIC-Datenbank (Medical Data Mart for Intensive Care) aus den USA. Die MIMIC-Datenbank stellt intensivmedizinische Routinedaten eines amerikanischen Maximalversorgers für Forschungszwecke zur Verfügung und wurde über einen langen Zeitraum iterativ entwickelt. Erfasst werden von MIMIC eine Vielzahl feingranularer Daten wie beispielsweise Vitalparameter, Medikationen, Labormessungen, Diagnosecodes, Überlebensdaten. In sogenannter „de-identifizierter" Form werden diese Daten im Rahmen einer Nutzungsvereinbarung weltweit für Forschungszwecke zugänglich gemacht. Durch ihren Open-Data-Ansatz ermöglicht es MIMIC, bereits vorhandene klinische Daten für ein breites Spektrum an weiteren wissenschaftlichen Forschungszwecken nutzbar zu machen.

In einem ersten Schritt wird MIMIC in Bezug auf die Datenverfügbarkeit und rechtliche Rahmenbedingungen beschrieben und der Aspekt der Repräsentativität und Übertragbarkeit auf Deutschland betrachtet.

Im zweiten Schritt wird untersucht, weshalb es bisher kein deutsches Pendant der MIMIC gibt. Als wesentliche Ursachen hierfür werden die strukturelle Organisation des deutschen Gesundheitswesens sowie die Herausforderungen und Rechtsunsicherheiten in Bezug auf den Datenschutz herausgearbeitet.

Im dritten Schritt wird untersucht, inwieweit ein deutsches Pendant zur MIMIC einen Mehrwert schaffen würde. Es wird festgestellt, dass eine einfache Reproduktion von MIMIC möglicherweise nicht das optimale Vorgehen wäre. Stattdessen könnte eine ambitioniertere Herangehensweise sinnvoll sein, die die Repräsentativität der Daten verbessert, über die Intensivmedizin hinausgeht und auch moderne Datenarten einschließt. Dies könnte den Forschungsnutzen maximieren und gleichzeitig strukturelle Voraussetzungen schaffen, die die systematische Sekundärnutzung von Routinedaten als integralen Bestandteil eines kontinuierlichen, iterativen Verbesserungsprozesses der Versorgung von Patientinnen und Patienten im Sinne eines lernenden Gesundheitssystems etablieren helfen.

Im vierten Schritt werden Umsetzungsmöglichkeiten analysiert. Die Herausforderungen bei der Schaffung einer deutschen MIMIC umfassen rechtliche, technische und strukturelle Aspekte, wie etwa die Sicherstellung der Repräsentativität der Daten, den Umgang mit Datenschutzanforderungen und die Integration relevanter Datenquellen.

In Bezug auf den zentralen Aspekt des Datenschutzes werden Anonymisierungsverfahren für unterschiedliche medizinische Datenarten eingehend untersucht, da diese auch bei der ursprünglichen MIMIC eine zentrale Bedeutung haben. Es wird aufgezeigt, dass diese zwar einen starken Schutz bieten können, aber im Einsatz mit Rechtsunsicherheiten behaftet sind sowie die Nutzbarkeit der Daten einschränken können und daher mit weiteren technischen und organisatorischen Maßnahmen kombiniert werden sollten. Dafür werden unterschiedliche Umsetzungsszenarien vorgeschlagen und weiter untersucht. Diese umfassen eine Bereitstellung in sicheren Verarbeitungsumgebungen oder mittels föderierter Ansätze, was möglicherweise die Nützlichkeit für wissenschaftliche Zwecke einschränkt. Ein einwilligungsbasiertes Vorgehen könnte weitere Zugangsmöglichkeiten eröffnen, setzt jedoch eine umfangreiche Vorbereitungs- und Einwilligungsphase voraus und wirft Fragen zur Repräsentativität auf.

Es folgt eine Analyse der rechtlichen Rahmenbedingungen unter Berücksichtigung der Umsetzungsszenarien. Für die Schaffung einer "deutschen MIMIC" auf Basis von Datenanonymisierung ist eine rechtssichere Definition der Anonymität von Daten notwendig, während für Szenarien mit personenbezogenen Daten eine spezifische rechtliche Legitimationsgrundlage benötigt wird. Das Gesundheitsdatennutzungsgesetz (GDNG) schafft eine Grundlage für die Verarbeitung von Versorgungsdaten zu Forschungszwecken und unterstützt damit die Konzeption einer "deutschen MIMIC", die Daten anonymisiert aufbereitet. Für die Weitergabe personenbezogener Daten an Dritte sind jedoch weitere rechtliche Grundlagen erforderlich, da das GDNG hierfür keine Regelung bietet. Wesentliche Herausforderungen umfassen die Rechtszersplitterung und die daraus folgende Notwendigkeit einer Rechtsvereinheitlichung, um länderübergreifende Forschung zu erleichtern und die Nutzung von Daten zu Forschungszwecken rechtssicher zu gestalten.

Abschließend werden Herausforderungen für die Implementierung von Anonymität auf Datenebene nochmals detailliert beleuchtet und eine prozessorientierte Sicht auf Anonymität vorgeschlagen, die einen Weg darstellen könnte, praktikable und operationalisierbare Anforderungen an die Anonymisierung und den Datenschutz in Forschungsvorhaben zu etablieren. Dies könnte eine zeitnahe und ressourceneffiziente Umsetzung ermöglichen und die internationale Kompetitivität der deutschen intensivmedizinischen Forschung stärken.

A. Einleitung

Die Medical Data Mart for Intensive Care (MIMIC[1]) Datenbank ist ein wichtiger, in den Vereinigten Staaten von Amerika (USA) über Jahrzehnte iterativ weiterentwickelter und breit verfügbar gemachter Datensatz, der im Wesentlichen auf der systematischen Aufbereitung und Bereitstellung klinischer Routinedaten eines großen amerikanischen Maximalversorgers basiert. Das vorliegende Gutachten untersucht die Möglichkeit und Sinnhaftigkeit des Aufbaus eines deutschen Pendants zur amerikanischen MIMIC. Basierend auf den Erfahrungen mit der heute verfügbaren MIMIC und verwandten Initiativen wird das Spektrum möglicher Handlungsoptionen abgeleitet sowie der erforderliche Ressourceneinsatz und das Verhältnis zwischen Aufwand und Nutzen bewertet.

Insbesondere die Beobachtungen zur Nutzbarkeit und zukunftssicheren translationalen Wirksamkeit von MIMIC-analogen Datenbereitstellungsmaßnahmen machen deutlich, dass das Optimum des Aufwand-Nutzenverhältnisses der verfügbaren Handlungsoptionen möglicherweise bei einem Maßnahmenumfang liegt, der über die reine Reproduktion der MIMIC in Deutschland hinausgeht. Stattdessen könnte eine deutlich ambitioniertere, strategische Herangehensweise an den MIMIC-artigen Strukturaufbau erwägenswert sein. Auf Basis möglicher Maßnahmen, die sich auf Handlungsfelder wie klinische Datenharmonisierung in unterschiedlichen Prozessschritten, Aufbau einer rechtskonformen Akquisitions-, Bereitstellungs- und Analyseinfrastruktur und Maßnahmen zur Unterstützung der Schöpfung unmittelbaren systemischen Mehrwertes aus den aufgebauten Infrastrukturen beziehen, wird versucht, das

[1] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35. PMID: 27219127; PMCID: PMC4878278.

Spektrum möglicher Handlungsoptionen und resultierender Ressourcenbedarfe orientierend aufzuspannen und bzgl. möglicher Wirksamkeit und Mehrwerte einzuschätzen.

Das für die Intensivmedizin skizzierte Vorgehen kann dabei in vielerlei Hinsicht auch prototypisch für mögliche nächste Schritte bei der Etablierung der infrastrukturellen Voraussetzungen eines dynamisch datengetrieben lernenden, patientinnen- und patientenorientierten Gesundheitssystems verstanden werden. Die Intensivmedizin eignet sich hierfür aus vielerlei Gründen in besonderer Weise. Dies liegt zum einen am intrinsisch hohen Technologisierungsgrad und der überdurchschnittlichen Verfügbarkeit von auch technisch hochqualifiziertem und interessiertem medizinischen Fachpersonal. Ein weiterer wesentlicher Faktor ist der extrem hohe Ressourceneinsatz in der Versorgung der Patientinnen und Patienten, der die Gewinnschwelle für potentiell Qualität, Effizienz und Innovationsgeschwindigkeit verbessernde Maßnahmen in Richtung ambitionierterer Vorgehensweisen verschiebt und so etwaige Investitionen früher und besser rechtfertigen hilft.

B. Hintergrund: Was ist die MIMIC-Datenbank und was kann sie leisten?

I. Historische Entwicklung

Der Medical Data Mart for Intensive Care[2] (früher: Multiparameter Intelligent Monitoring in Intensive Care, MIMIC) als wichtiger Teil der breiter angelegten Physionet-Initiative[3] ist eine in den späten 90er Jahren des vergangenen Jahrhunderts konzeptionierte und in den frühen 2000ern als MIMIC-II in größerem Maßstab vorangetriebene Initiative, die die Aufbereitung und breite Verfügbarmachung möglichst vollständiger intensivmedizinischer Routinedaten unterschiedlicher Modalitäten zur Algorithmen- und Modellentwicklung und -validierung zum Ziel hat[4]. Ein wesentliches Ziel war hierbei von Anfang an die Berücksichtigung von hochauflösenden physiologischen Signalen wie Elektrokardiogramme (EKGs) und invasive gemessene Blutdrücke, also Biosignaldaten in voller gemessener Zeitauflösung von oft über 100 Hertz (d.h. Messwerten *pro Sekunde*), und deren nahtlose Verknüpfung mit den übrigen, niedriger aufgelösten elektronischen Dokumentationsdaten, die üblicherweise maximal einen Wert *pro Minute* enthalten. Zu Beginn der Umsetzung war dieses Ziel nur in enger Zusammenarbeit mit

[2] Moody GB, Mark RG. (1996) A database to support development and evaluation of intelligent intensive care monitoring. In: Computers in Cardiology; 1996 Sep; 657-660. IEEE.

[3] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, u. a. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 13. Juni 2000;101(23):E215–20.

[4] Die ursprüngliche Konzeption geht wesentlich auf das Laboratory for Computational Physiology am Massachusetts Institute of Technology (MIT) in Kooperation mit lokalen akademischen (Margret and H.A. Rey Institute for Nonlinear Dynamics) und klinischen Partnern (Beth Israel Deaconess Medical Center [BIDMC], beide Teil von bzw. assoziiert mit der Harvard Medical School) und Industriepartnern (Philips Medical Systems), zurück und wurde in der Folgezeit prototypisch u.a. mit Unterstützung der National Aeronautics and Space Administration (NASA) etabliert sowie durch die National Institutes of Health (NIH) unterstützt; vgl. Mark R. The Story of MIMIC. In: MIT Critical Data, Herausgeber. Secondary Analysis of Electronic Health Records [Internet]. Cham: Springer International Publishing; 2016 [zitiert 3. Dezember 2023]. S. 43–9.

dem Hersteller der Monitoringsysteme und dann auch nur unvollständig erreichbar, da die damals übliche Technologie keine niederschwelligen Mechanismen zur Extraktion hochauflösender Rohdaten aus den klinisch genutzten Systemen vorsah.

Diese hochauflösenden Signale wurden der Nutzergemeinschaft erstmalig als sog. “MIMIC II Waveform Database” zur Verfügung gestellt und ergänzten die übrigen strukturierten Dokumentationsdaten als Teil von MIMIC II.[5] Eine weitere wichtige Datenquelle für die MIMIC-Initiative sind extern erhobene Mortalitätsdaten, die nach Zusammenführung mit den Krankenhausdaten eine Nachverfolgung des MIMIC-Kollektivs bzgl. des intensivmedizinisch hochrelevanten 28-Tage- und 1-Jahres-Überlebens auch nach Ende der Krankenhausbehandlung bei Patientinnen und Patienten, die lebend entlassen werden, erlauben. Sämtliche Daten wurden entsprechend der Vorgaben des Health Insurance Portability and Accountability Act (HIPAA) de-identifiziert und zusätzlich durch Veränderung der Zeitstempel unter weitestmöglicher Bewahrung potenziell medizinisch relevanter Aspekte wie Saisonalität gegen Re-Identifizierungsversuche geschützt, wobei eine Kombination aus automatisierten Verfahren sowie manueller Prozesse zur Entfernung identifizierender Merkmale und zur Qualitätssicherung zum Einsatz kam.[6] MIMIC-II-Biosignaldaten wurden vollständig öffentlich zur Verfügung gestellt, während die klinischen Daten aus MIMIC-II Forschenden nach Nachweis der erfolgreichen Absolvierung eines u.a. die HIPAA-Regularien zum Datenschutz behandelnden Kurses und Abschluss einer Datennutzungsvereinbarung zugänglich gemacht wurden.

Nachdem sich MIMIC und MIMIC-II bereits als außerordentlich hilfreich und wichtig für die Forschungsgemeinschaft erwiesen hatten, wurde das Spektrum der erfassten Datenarten wie auch der Umfang der erfassten Behandlungsepisoden in dem Folgeprojekt MIMIC-III[7] nochmals erheblich erweitert. Erfasst wurden von MIMIC III u.a. eine größere Anzahl von Patientinnen und Patienten (über 53000 Krankenhausaufnahmen in MIMIC-III Clinical vs. etwas über 25000 Intensivaufenthalte in MIMIC-II Clinical[8]) und eine breitere Abdeckung von medizinischen Domänen in einem sich mit dem Erfassungszeitraum von MIMIC-II überlappenden, aber diesen erweiternden Zeitraum. Im aktuellen Projekt MIMIC-IV[9] schließlich wurde MIMIC-III nochmal erweitert und restrukturiert. Der Gesamtdatensatz ist nun modular gegliedert, das Datenbankschema wurde bereinigt, um die Sekundärnutzung weiter zu erleichtern, und die Dokumentation der Datenherkunft und -entstehung findet stärkere Beachtung. Weiterhin wurde die Anreicherung des Datenbestandes um Mortalitätsinformationen aufgrund von Repräsentativitätsproblemen des sogenannten Death Master File der United States Social

[5] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In: Computers in Cardiology; 2002 Sep; 641-644. IEEE.
[6] Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-assisted de-identification of free text in the MIMIC II database. In: Computers in Cardiology, 2004; 2004 Sep; 341-344. IEEE.
[7] Johnson AEW, Pollard TJ, Shen L, Lehman L wei H, Feng M, Ghassemi M, u. a. MIMIC-III, a freely accessible critical care database. Sci Data. 24. Mai 2016;3(1):160035.
[8] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, u. a. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Crit Care Med. Mai 2011;39(5):952–60.
[9] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, u. a. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 3. Januar 2023;10(1):1.

Security Administration[10] auf das Registry of Vital Records and Statistics des Bundesstaates Massachusetts umgestellt.

II. Datenarten und Umfang

MIMIC hat es sich zum Ziel gesetzt, “Reale-Welt-Daten” (Real World Data, RWD), d.h. also möglichst unverfälschte, nicht bereinigte oder sonst wie veränderte Daten über Diagnostik und Behandlung von Intensivpatientinnen und -patienten sowohl aus der elektronischen Dokumentation als auch direkt aus den zur Überwachung, Diagnostik und Therapie eingesetzten Medizingeräten zur Ermöglichung von medizinisch relevanter Forschung und Algorithmenentwicklung breit verfügbar zu machen.[11] Dieser Ansatz unterscheidet sich fundamental von klassischen Kohortenstudien, die prospektiv den Einschlusskriterien entsprechende Probandinnen und Probanden rekrutieren und dann eine studienspezifische, meist aufwändig qualitätsgesicherte Datenerhebung durchführen. Er unterscheidet sich auch von medizinischen Registern, die zwar versuchen, bestimmten Selektionskriterien (z.B. der Durchführung definierter medizinischer Eingriffe) genügende Populationen möglichst vollständig oder zumindest repräsentativ zu erfassen, aber anders als MIMIC eine meist manuelle, zum Versorgungsprozess additive Datenerhebung durchführen. Diese können sich durchaus auch zumindest teilweise aus RWD speisen[12].

Die MIMIC-Population umfasste zum Zeitpunkt der Veröffentlichung von MIMIC-IV[13] eine ad-hoc-Stichprobe von zwischen 2008 und 2019 am Beth Israel Deaconess Medical Center[14] in der Notaufnahme oder Intensivstation behandelten Patientinnen und Patienten, unter Ausschluss von solchen mit einem Lebensalter unter 18 Jahren zum Zeitpunkt des Erstkontaktes oder mit besonderem Schutzbedarf. Die Daten in MIMIC umfassen[15]:

- **Demographie, Administration, Mortalität:** In dieser Kategorie werden wesentliche demografische Daten und administrative Informationen, darunter Aufnahme- und Entlassungsdaten der Patientinnen und Patienten sowie Todeszeitpunkte, Aufnahmegrund und der Aufnahmeart erfasst.

[10] Levin MA, Lin HM, Prabhakar G, McCormick PJ, Egorova NN. Alive or dead: Validity of the Social Security Administration Death Master File after 2011. Health Services Research. 2019;54(1):24–33.
[11] Sämtliche folgenden Ausführungen beziehen sich auf das zum Zeitpunkt der Erstellung dieses Textes aktuelle MIMIC-IV, welches allerdings einer kontinuierlichen Weiterentwicklung unterliegt, so dass in entscheidungsrelevanten Szenarien ein Abgleich mit dem aktuellen Stand an der Originalquelle (https://www.physionet.org) empfohlen wird.
[12] Gutachten zur Weiterentwicklung medizinischer Register zur Verbesserung der Dateneinspeisung und -anschlussfähigkeit [Internet]. 2021 [zitiert 3. Dezember 2023]. Verfügbar unter: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Gesundheit/Berichte/REG-GUT-2021_Registergutachten_BQS-TMF-Gutachtenteam_2021-10-29.pdf
[13] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, u. a. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 3. Januar 2023;10(1):1.
[14] Beth Israel Deaconess Medical Center
[15] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215–e220 sowie Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035.

- **Klinische Informationen:** Diese Kategorie beinhaltet wichtige klinische Daten wie Vitalzeichen, Laborergebnisse, Diagnosen, durchgeführte Prozeduren sowie Medikation.

- **Daten aus bildgebenden Verfahren:** In dieser Kategorie sind Radiologieberichte und in einigen Versionen Verknüpfungen zu bildgebenden Untersuchungen enthalten.

- **Physiologische Überwachungsdaten:** Diese Kategorie umfasst EKG-Aufzeichnungen und andere relevante Überwachungsdaten wie z.B. invasiv gemessene Blutdruckkurven.

III. Rechtsgrundlagen

Die rechtliche Zulässigkeit der MIMIC-Datenbanken fußt darauf, dass die verarbeiteten Daten der Patientinnen und Patienten nach den Maßstäben der HIPAA Privacy Rule als "de-identifiziert" eingeordnet werden. Nach § 164.514 HIPAA Privacy Rule ist von solcherlei de-identifizierten Daten dann auszugehen, wenn diese Daten nicht mehr zur Identifizierung einer Person dienen und es auch "keine vernünftige Grundlage für die Annahme gibt", dass diese Daten noch zur Identifizierung einer Person verwendet werden könnten (*"no **reasonable basis to believe** that the information can be used to identify an individual"*). Das Grundverständnis entspricht damit im Wesentlichen dem, was auch nach europäischem Verständnis anonymisierte Daten kennzeichnet. Auch die Datenschutz-Grundverordnung (DS-GVO) stellt für die Frage der Identifizierbarkeit einer Person auf das „allgemeine Ermessen" ab: Zu berücksichtigen sind nach Erwägungsgrund (EG) 26 der DS-GVO alle Erkenntnismittel, die nach allgemeinem Ermessen wahrscheinlich zur Identifizierung genutzt werden *("account should be taken of all the **means reasonably likely to be used**... to identify the natural person directly or indirectly").*[16] Damit einher geht dann allerdings der wesentliche Unterschied, dass sich in HIPAA weitere, spezifische Vorgaben zur „De-Identifizierung" von Daten finden, die eine verlässliche und rechtssichere Differenzierung zwischen personenbezogenen ("individually identifiable") und anonymisierten ("de-identified") Gesundheitsdaten ermöglichen.

Grundsätzlich kommen nach der HIPAA Privacy Rule zwei Methoden einer De-Identifizierung von Gesundheitsdaten in Betracht.[17] Zum einen kann die De-Identifizierung durch qualifizierte Expertinnen oder Experten formal festgestellt werden (§ 164.514(b)(1) HIPAA Privacy Rule). Zum anderen kann die De-Identifizierung von Gesundheitsdaten nach der "Safe Harbor"-Methode unter der Voraussetzung angenommen werden, dass eine Reihe von Identifikatoren aus dem Datenbestand entfernt werden sowie die jeweilige verantwortliche Stelle keine tat-

[16] Finck M, Pallas F, They who must not be identified—distinguishing personal from non-personal data under the GDPR, *International Data Privacy Law*, Volume 10, Issue 1, February 2020, Pages 11–36: " The test devised by Recital 26 GDPR essentially embraces a risk-based approach to qualify information. Where there is a reasonable risk of identification, data ought to be treated as personal data. Where that risk is merely negligent, data can be treated as non-personal data, and this even though identification cannot be excluded with absolute certainty."

[17] HHS. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [Internet]. 2022 [zitiert 20. Dezember 2023]. Verfügbar unter: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale.

sächliche Kenntnis davon hat, dass die verbleibenden Informationen allein oder in Kombination mit anderen Informationen zur Identifizierung einer bestimmten Person verwendet werden könnten (§ 164.514(b)(2) HIPAA Privacy Rule).

Im Fall der MIMIC stützen sich die Verantwortlichen auf letztere Alternative einer De-Identifizierung, indem sie sämtliche in der "Safe Harbor"-Methode aufgelisteten Identifikatoren aus dem Datenbestand entfernen.[18] Die Liste umfasst insgesamt 18 Punkte, wovon die ersten 17 sich explizit auf spezifische Angaben beziehen und der letzte Punkt jedes weitere Merkmal, das eine eindeutige Identifizierung von Personen ermöglicht, umfasst. Zu den explizit genannten Merkmalen gehören u.a. Namen, bestimmte Adress- und Zeitangaben oder Telefon-, Faxnummer und Mailadresse, Sozialversicherungs- und Kontonummern sowie Gerätekennungen und Seriennummern, Fotografien des gesamten Gesichts und vergleichbare Bilder.

IV. Grenzen und Perspektiven der heutigen MIMIC

1. Repräsentativität der Population als zentrale Herausforderung

Für die Entwicklung datengetriebener Erkenntnisse und Innovationen unter Nutzung großer Datenmengen und maschineller Lernverfahren ist eine repräsentative Abbildung sowohl der Population der Patientinnen und Patienten als auch der Versorgungspraxis entscheidend, um systematische Benachteiligungen von Patientinnen und Patienten sowie potentiell gefährliche Fehlinterpretationen aufgrund unpassender Kontextualisierung von Entscheidungen zu vermeiden.

Was die repräsentative Abbildung der relevanten Population von Patientinnen und Patienten angeht, ist im Fall der MIMIC problematisch, dass es sich hierbei um einen im Wesentlichen monozentrisch erhobenen Datensatz handelt. Selbst für die Population der USA kann dieser Datensatz daher aus zwei Gründen nicht vollständig repräsentativ sein: Zum einen weist auch die USA eine z.B. in ethnischer Zusammensetzung örtlich nicht vollständig homogene Zusammensetzung der Bevölkerung auf und somit kann der Einzugsbereich der datenliefernden Kliniken nicht vollständig repräsentativ für die US-Bevölkerung sein. Zum anderen ist davon auszugehen, dass auch die Art der Institution selbst (international führender, klinisch und akademisch extrem sichtbarer Maximalversorger) eine Selektionsverzerrung im Patientinnen- und Patientenkollektiv induziert.

Bezüglich der Repräsentativität der MIMIC-Daten für die Versorgungspraxis bestehen potentiell noch größere Herausforderungen, da die Daten nur die Versorgungspraxis an einem akademisch führenden Maximalversorger reflektieren können, die üblichen Vorgehensweisen in Häusern geringerer Versorgungsstufe und mit geringerer Ressourcenausstattung hingegen im Datensatz nicht adäquat repräsentiert sein können. Zudem ist auch zu erwarten, dass es durch die gegenüber der flächendeckenden Patientinnen- und Patientenversorgung insgesamt sicherlich gegebene Überrepräsentation innovativer und forschungsgetriebener oder -

[18] Johnson AEW, Pollard TJ, Shen L, Lehman L wei H, Feng M, Ghassemi M, u. a. MIMIC-III, a freely accessible critical care database. Sci Data. 24. Mai 2016;3(1):160035; Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, u. a. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 3. Januar 2023;10(1):1.

begleiteter Handlungsweisen in einem solchen Zentrum zu Verzerrungen der Datenbestände kommt, die die Generalisierbarkeit von aus diesen Datenbeständen abgeleiteten Schlussfolgerungen und Algorithmen auf die intensivmedizinische Versorgung *insgesamt* in Frage stellen.[19]

Da die Heterogenität der regionalen Patientinnen und Patientenpopulationen und Versorgungsstrukturen in Deutschland ebenfalls signifikant ist, wird eine zukunftssichere deutsche MIMIC die Repräsentativitätsfragen bzgl. Eigenschaften der Patientinnen und Patienten sowie Versorgungskontexten proaktiv und systematisch adressieren müssen, idealerweise durch ein skalierbares, multizentrisches Vorgehensmodell.

2. Datenabdeckung

Bzgl. der inhaltlichen Abdeckung der aktuellen MIMIC-Variante zeigen sich einige zumindest perspektivisch immer relevantere Abdeckungslücken, die in einer neu aufzubauenden deutschen MIMIC zumindest in der mittelfristigen Planung aktiv adressiert werden könnten und sollten:

- Diverse klinisch relevante Informationen liegen nicht strukturiert, sondern in Form nicht standardisierter Freitexte vor, was die Vergleichbarkeit reduziert und eine verlässliche Nutzung in größer angelegten Analyse- und Entwicklungskontexten erschwert; darüber hinaus sind unstrukturierte Eingabedaten auch als Datenquelle für sicherheitskritische Entscheidungsunterstützungsverfahren problematisch.

- Biosignal- und Bilddaten liegen nur für einen Teil des Gesamtdatensatzes und/oder unvollständig vor.

- Schon aufgrund ganz anders gelagerter Re-Identifizierungsrisiken kann MIMIC nicht ohne weiteres auch die Ergebnisse genetischer Analysen und anderer OMICS-Verfahren, also Verfahren zur Analyse der Strukturen, Funktionen und Dynamiken verschiedener Arten biologischer Systeme auf molekularer Ebene, einschließen.[20]

- Abseits von Mortalitätsraten stellt MIMIC bisher nur wenige für die Ergebnisqualitätsmessung wichtige von Patientinnen und Patienten berichtete Informationen zu Behandlungsergebnissen zur Verfügung. Eine harmonisierte Herangehensweise an diesen Themenkomplex wird angesichts der wachsenden Evidenz zu den erheblichen

[19] Die Treiber hinter MIMIC haben u.a. mit der Aufbereitung und Bereitstellung des eICU-CRD-Datensatzes bereits selbst begonnen, diese Herausforderungen durch eine auf telemedizinischen Infrastrukturen aufsetzende multizentrische Herangehensweise zu adressieren (REF); eICU-CRD deckt bereits über 200 Krankenhäuser mit über 300 Intensivstationen aus weiten Teilen der USA ab und adressiert somit bereits einige der o.g. Herausforderungen für die US-amerikanische Patientenpopulation zumindest partiell, indem hier ein völlig anderes Versorgungspektrum abgedeckt wird. Repräsentativ ist aber auch dieser Datensatz nicht; ausführlicher dazu O'Halloran HM, Kwong K, Veldhoen RA, Maslove DM. Characterizing the Patients, Hospitals, and Data Quality of the eICU Collaborative Research Database*. Critical Care Medicine. Dezember 2020;48(12):1737.

[20] Martinez C, Jonker E. A Practical Path Toward Genetic Privacy in the United States. Verfügbar unter: https://fpf.org/wp-content/uploads/2020/04/ APracticalPathTowardGeneticPrivacy_April2020.pdf.

langfristigen Auswirkungen der Intensivtherapie auf die überlebenden Patientinnen und Patienten immer wichtiger[21].

Die beschriebenen Abdeckungslücken sind in Teilen schlicht der Genese von MIMIC als reinem Routinedatensatz geschuldet, in dem klinische Aspekte nach heutigem Stand der Technik regelmäßig eben nicht in standardisierter und strukturierter Weise erfasst werden. Für die Zukunft lassen sich solche Lücken plausibel nur durch eine Weiterentwicklung der routinemäßigen Behandlungsdokumentation systematisch adressieren.

3. Übertragbarkeit auf und Vergleichbarkeit mit deutschen Verhältnissen

Ist MIMIC schon bzgl. der Population und Behandlungspraxis in den USA nur eingeschränkt repräsentativ, so gilt dies erst recht für die Repräsentation von und damit Übertragbarkeit auf europäische – und im Speziellen deutsche – Verhältnisse. Ein wichtiger Aspekt sind die offensichtlichen Abweichungen in der ethnischen Zusammensetzung der Patientinnen- und Patientenpopulation und der Einfluss der abweichenden Lebensverhältnisse und Gesellschaftsstrukturen. Des Weiteren spielt vor allem auch die variable Organisation der intensivmedizinischen Behandlung eine Rolle, in der die Arbeitsverteilung zwischen ärztlichen und nicht-ärztlichen Berufsgruppen z.T. erheblich divergiert. Auch die substantiell unterschiedliche Handhabung der intensivmedizinischen Ressourcenallokation, , beispielsweise betreffend die Intensivbettendichte, ist relevant, was den Vergleich und Transfer internationaler Forschungsergebnisse herausfordernd gestaltet.[22]

C. Machbarkeit einer deutschen MIMIC

I. Ursachenanalyse: Warum gibt es bislang keine deutsche MIMIC?

Anders als bspw. in den Niederlanden, wo die AmsterdamUMCdb[23] aufgebaut wurde, gibt es bis dato in Deutschland kein Äquivalent zu einer MIMIC oder einer vergleichbar konzipierten Datenbank. Geschuldet ist diese eingeschränkte Verfügbarkeit feingranularer intensivmedizinischer Routinedaten (aber auch feingranularer klinischer Daten aus anderen Bereichen der Medizin) hierzulande einer Kombination aus technischen, rechtlichen und strukturellen Gegebenheiten, die Deutschland trotz international eigentlich durchaus kompetitiver Forschungs-

[21] Turnbull AE, Rabiee A, Davis WE, Nasser MF, Venna VR, Lolitha R et al. Outcome Measurement in ICU Survivorship Research from 1970-2013: A Scoping Review of 425 Publications. Crit Care Med. Juli 2016;44(7):1267–77.

[22] Murthy S, Wunsch H. Clinical review: International comparisons in critical care - lessons learned. Critical Care. 5. April 2012;16(2):218.

[23] Thoral, P. J., Peppink, J. M., Driessen, R. H., Sijbrands, E. J. G., Kompanje, E. J. O., Kaplan, L., Bailey, H., Kesecioglu, J., Cecconi, M., Churpek, M., Clermont, G., van der Schaar, M., Ercole, A., Girbes, A. R. J., Elbers, P. W. G., on behalf of the Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force (2021). Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med. 2021 Jun 1;49(6):e563-e577

und Versorgungsstrukturen in vielen Innovationsfeldern letztlich von Datenquellen aus anderen Ländern abhängig machen.

1. Grundsätzliche Probleme

Ursächlich für diese Situation sind u.a. folgende Aspekte, die zum Teil im Sinne einer gegenseitigen, mehr als additiven Verstärkung zusammenwirken:

Die **langjährige Unterfinanzierung** der Leistungserbringung im Bereich technischer Investitionen, insbesondere im Bereich der Informationstechnologie, sowie für den Unterhalt und Betrieb erforderlicher Strukturen führt dazu, dass schon die Datenquellen für eine mögliche MIMIC in Form ausgereifter elektronischer Dokumentationssysteme an vielen Häusern nicht oder nur rudimentär vorhanden sind. Ebenso fehlt es an einer dauerhaften Speicherinfrastruktur für hochauflösende physiologische Messdaten, die i.d.R. nicht als Teil der elektronischen Dokumentation aufgezeichnet werden, aber für Qualitätssicherung und Forschung potentiell hochrelevant sind.

Wesentliche Hürde für jede Form der gemeinsamen Datennutzung zu Forschungszwecken ist auch die **inter- und transsektoral zersplitterte strukturelle Aufstellung** des deutschen Gesundheitssystems. Es liegt nicht nur eine Zergliederung der Versorgungslandschaft in stationäre und ambulante Leistungserbringer mit völlig unterschiedlichen und nicht tief miteinander verzahnten Strukturen vor, sondern auch noch eine hierzu orthogonale Teilung der Finanzierungsmodelle (und damit wesentlicher Teile der leistungserbringerübergreifenden Organisations- und technischen Unterstützungsstrukturen) in den Bereich der gesetzlichen Krankenversicherungen und privaten Krankenversicherungen.

Eine weitere spezifisch deutsche Problematik im Bereich der Forschungsdatenverarbeitung ist die der **Rechtszersplitterung und Rechtsunsicherheit**. Die föderale Zergliederung der Verwaltungsstrukturen und zugehörigen Rechtsnormen erschweren die rechtssichere Etablierung einer deutschen MIMIC ganz erheblich, weil je nachdem, welche Organisationsform die beteiligten Institutionen haben – Einrichtung in privater Trägerschaft, Einrichtungen des Bundes, der Länder oder auch der Kirchen – und in welchem Bundesland sie angesiedelt sind, jeweils andere Rechtsvorgaben zu beachten sind.

2. Restriktives Forschungsdatenschutzrecht

Zentrales Hemmnis für die Forschungsdatenverarbeitung ist in Deutschland der teils noch immer **zu restriktive Ansatz** des Datenschutzrechts hinsichtlich der Zulässigkeit einer Datenverarbeitung zu Forschungszwecken.

a) (Forschungsfreundlicher) Ausgangspunkt: DS-GVO und § 27 BDSG

An sich ist spätestens mit Geltung der DS-GVO auf europäischer Ebene das Datenschutzrecht durch einen ausgesprochen forschungsfreundlichen Ansatz geprägt und das Beispiel der Niederlande mit der AmsterdamUMCdb zeigt, dass unter Geltung der DS-GVO ein Modell wie die MIMIC rechtlich durchaus umsetzbar wäre.

Die forschungsfreundliche Grundausrichtung des Datenschutzrechts setzt sich dann zunächst auch einmal im deutschen Recht auf Ebene des Bundesdatenschutzgesetzes (BDSG) fort. So sieht § 27 BDSG, in Umsetzung der Öffnungsklausel des Art. 9 Abs. 2 lit. j DS-GVO, die Zulässigkeit einer Datenverarbeitung zu Forschungszwecken auch ohne Einwilligung vor, „wenn die Verarbeitung zu diesen Zwecken erforderlich ist und die Interessen des Verantwortlichen an der Verarbeitung die Interessen der betroffenen Person an einem Ausschluss der Verarbeitung erheblich überwiegen." Grundidee sowohl der DS-GVO als auch des § 27 BDSG ist, dass eine Datenverarbeitung zu Forschungszwecken im Rahmen der Erforderlichkeit grundsätzlich zulässig sein soll.[24] Ergänzt wird dieser Erforderlichkeitsgrundsatz durch das Gebot einer Interessenabwägung zwischen Forschungsinteressen einerseits und Vertraulichkeitsinteressen der betroffenen Person andererseits, um damit dem Verhältnismäßigkeitsgebot Rechnung zu tragen.

Forschungsfreundlich ist dieses Grundkonzept von DS-GVO und BDSG vor allem auch deshalb, weil es im Rahmen der Interessenabwägung entscheidend darauf abstellt, ob und inwieweit Verantwortliche mittels Garantien und Schutzmaßnahmen die Vertraulichkeits- und Datenschutzinteressen der betroffenen Personen wahren.[25] Forschende haben es damit selbst in der Hand, die Interessenabwägung zu ihren Gunsten zu beeinflussen, indem sie durch entsprechende technische oder organisatorische Maßnahmen einen bestmöglichen Datenschutz zugunsten der betroffenen Personen gewährleisten. [24]

b) Anwendungsprobleme in der Praxis

Zu konstatieren ist allerdings, dass das auf dem Papier forschungsfreundlich ausgestaltete Grundkonzept von DS-GVO und BDSG in der Praxis bislang nicht angekommen ist und damit die vorhandenen rechtlichen Handlungsspielräume nicht ausgeschöpft werden (können). Zurückzuführen ist dies zunächst einmal darauf, dass besagte Interessenabwägung gerade in Deutschland von einer (partiell auch historisch mitbedingten) kulturellen, aber auch rechtspolitischen Tendenz geprägt ist, die gesellschaftliche Diskussion und notwendige Abwägungen zur Datennutzung stark an den Risiken und weniger an den Chancen und den ebenfalls potentiell schädlichen Folgen einer Nicht-Nutzung zu orientieren. Hinzu tritt das Problem einer fehlenden praxisgerechten Operationalisierung der geforderten Interessenabwägung. Forschende sehen sich der Herausforderung gegenüber, ex ante eine bestimmte Interessenabwägung vornehmen zu müssen, ohne rechtssicher absehen zu können, ob deren Ergebnis ex post möglicherweise durch Aufsichtsbehörden oder Gerichte wieder revidiert wird. Kritisiert wird, dass gerade in Deutschland die Datennutzung zu Forschungszwecken erheblich erschwert wird, zum einen durch eine gleichermaßen restriktive wie heterogene Gesetzgebung,

[24] Buchner, B. Forschungsdaten effektiver nutzen. Datenschutz Datensich 46, 555–560 (2022).
[25] Zur forschungsfreundlichen Ausrichtung der DS-GVO s. etwa Buchner B, Haber AC, Hahn HK, Prasser F, Kusch H, Sax U, Schmidt CO. Das Modell der Datentreuhand in der medizinischen Forschung. Datenschutz Datensicherheit-DuD. 2021;45:806-810; Buchner, B. Forschungsdaten effektiver nutzen. Datenschutz Datensich 46, 555–560 (2022); Spitz, M., Cornelius, K., Jungkunz, M. *et al.* Rechtlicher Rahmen für eine privilegierte Nutzung klinischer Daten zu Forschungszwecken . *MedR* **39**, 499–504 (2021) und nicht zuletzt auch der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit Kelber U. Wissenschaftliche Forschung – selbstverständlich mit Datenschutz, in: Roßnagel/Wallmann (Hrsg.). Stärkung der Forschung durch Datenschutz. Nomos 2023.

zum anderen aber auch durch eine uneinheitliche Auslegungspraxis seitens der Aufsichtsbehörden der Bundesländer und des Bundes.[26] Letztlich führen die restriktiven Rahmenbedingungen dazu, dass konkrete Gesundheitsfragestellungen, deren Beforschung an sich unstreitig im öffentlichen Gesundheitsinteresse liegt, hierzulande nicht bearbeitet werden können, weil sich die Forschungspraxis nicht mit den jeweiligen datenschutzrechtlichen Auflagen, sei es von Seiten der Datenschutzaufsicht, sei es von Seiten der internen Datenschutzkontrolle, vereinbaren lässt.[27] Abgeholfen werden könnte besagter Rechtsunsicherheit etwa durch klare Regeln dahingehend, dass eine bestimmte Behörde innerhalb einer bestimmten Frist über das Ergebnis einer Interessenabwägung entscheiden muss. Erste Ansätze in dieser Richtung finden sich im neuen Gesundheitsdatennutzungsgesetz (GDNG), wenn es um die Zulässigkeit einer gemeinsamen Datenverarbeitung innerhalb von Forschungsverbünden geht. § 6 Abs. 3 S. 4 GDNG setzt hier für die Zulässigkeit einer solchen gemeinsamen Datenverarbeitung u.a. die Zustimmung der zuständigen Datenschutzbehörde voraus, über die die Behörde "innerhalb eines Monats" entscheiden soll. Abzuwarten bleibt, welche Auslegung diese Regelung in der künftigen Praxis erfahren wird und ob der Zustimmung möglicherweise auch eine Bindungswirkung dahingehend zukommt, dass damit auch die anderen in § 6 Abs. 3 S. 4 GDNG normierten Zulässigkeitsvoraussetzungen als erfüllt anzusehen sind. Bestenfalls könnte dem Ansatz dann auch für andere Regelungsbereiche ein Modellcharakter zukommen, in denen das Ergebnis einer Interessenabwägung für die Nutzung existierender oder zukünftiger Erlaubnistatbestände von entscheidender Bedeutung ist, diese Abwägung aber durch an der Datenverarbeitung beteiligte Parteien kaum sinnvoll und vor allem nicht rechtssicher selbst getroffen werden kann.

c) Problem der Rechtszersplitterung auf landesrechtlicher Ebene

Nochmals erheblich forciert werden die Rechtsunsicherheit und die damit einhergehenden rechtlichen Hemmnisse für eine Forschungsdatenverarbeitung durch die Rechtszersplitterung auf landesrechtlicher Ebene. So ist es schon im Rahmen der allgemeinen Landesdatenschutzgesetze nur wenig zielführend, dass je nach Bundesland die Kriterien für eine Interessenabwägung zur Legitimation einer Forschungsdatenverarbeitung anders gefasst sind. Warum etwa nach § 17 Datenschutzgesetz Nordrhein-Westfalen (DSG NRW) eine Forschungsdatenverarbeitung zulässig sein soll, wenn schutzwürdige Belange der betroffenen Person „nicht überwiegen", nach § 13 Niedersächsisches Datenschutzgesetz (NDSG) demgegenüber bei der Interessenabwägung darauf abzustellen ist, ob ein schutzwürdiges Interesse der betroffenen Person „nicht entgegensteht" oder aber das öffentliche Interesse an der Durchführung des Forschungsvorhabens das schutzwürdige Interesse der betroffenen Person „überwiegt" oder nach § 18 Bremisches Ausführungsgesetz zur EU-Datenschutz-Grundverordnung (BremDSGVOAG) wiederum eine Datenverarbeitung nur erlaubt sein soll, soweit die Interessen auf Forschendenseite „erheblich überwiegen", lässt sich sinnvoll nicht begründen, sondern scheint vielmehr den individuellen Präferenzen des jeweiligen Landesgesetzgebers geschuldet zu sein. Selbst wenn jedes Bundesland dabei der festen Überzeugung sein sollte, es selbst am besten zu wissen, sollte für die Landesgesetzgeber gleichwohl offensichtlich sein,

[26] S. etwa die Kritik von Welzel, C; Cotte, F; Brückner, S; Lauber-Rönsberg, A; Muck, J; Gilbert, S. Gesundheitsdaten: Paradigmenwechsel steht auch in Deutschland bevor. Dtsch Arztebl 2023; 120(26): A 1162–5.

[27] Vgl. dazu etwa die konkreten Beispiele bei Merzenich, H; Sandkämper, L; Bäcker, M; Zeeb, H; Wollschläger, D. Datenschutzrecht: Wenn Mutlosigkeit zum Forschungshemmnis wird. Dtsch Arztebl 2023; 120(17): A761–4.

dass mit einer solchen Rechtszerfaserung niemandem gedient ist und die sinnvollen Anliegen medizinischer Forschung unnötig behindert werden.

Nochmals problematischer als im allgemeinen Landesdatenschutzrecht ist der Stand der Regulierung einer Forschungsdatenverarbeitung im Landeskrankenhausrecht: Ob und unter welchen Voraussetzungen Krankenhäuser Daten von Patientinnen und Patienten zu Forschungszwecken auch gegenüber Dritten offenlegen dürfen, ist in den einschlägigen Regelungen der Krankenhausgesetze der einzelnen Bundesländer völlig disparat geregelt und einer Vielzahl von ganz unterschiedlichen Einschränkungen unterworfen:

In Baden-Württemberg und Bayern etwa kommt eine Offenlegung von Daten von Patientinnen und Patienten überhaupt nur dann in Betracht, wenn das Krankenhaus eigene Forschungszwecke verfolgt.[28] Zusätzlich ist nach Art. 27 Bayerisches Krankenhausgesetz (BayKrG) Bedingung, dass die Daten in jedem Fall „im Gewahrsam des Krankenhauses verbleiben".[29] Großzügiger ist demgegenüber zwar etwa die Forschungsklausel im Berliner Landeskrankenhausrecht, zugleich ist diese Regelung allerdings überkomplex und kaum rechtssicher zu handhaben, wenn hier die unterschiedlichsten Varianten von Interessenabwägungsklauseln normiert werden, abhängig etwa davon, ob es um „eigene wissenschaftliche Forschungsvorhaben" geht oder ob es „zumutbar" ist, eine Einwilligung einzuholen.[30] In unmittelbarer Nachbarschaft, in Brandenburg, sehen sich Krankenhäuser wiederum einem gänzlich anderen Rechtskonzept gegenüber, wenn es hier für eine Offenlegung von Daten von Patientinnen und Patienten zu Forschungszwecken der vorherigen Bestätigung der zuständigen Rechtsaufsichtsbehörde sowie einer Anhörung der Landesdatenschutzbeauftragten bedarf.[31] In Schleswig-Holstein schließlich ist eine Forschungsdatenverarbeitung ohne Einwilligung der betroffenen Patientinnen und Patienten überhaupt nicht zulässig.[32]

Schon dieser kursorische Überblick macht deutlich, dass es bis dato vor allem mit Blick auf das Landesdatenschutzrecht an einem praktikablen und rechtssicheren Rahmen für die Datenverarbeitung zu Forschungszwecken fehlt. Insbesondere größer angelegte Forschungs- und Qualitätssicherungsvorhaben, an denen Gesundheitseinrichtungen unterschiedlicher Träger (Bund, Länder, Kirche, privatrechtlich) aus unterschiedlichen Bundesländern beteiligt sind, wie sie auch für eine deutsche MIMIC kennzeichnend wären, sehen sich regelmäßig mit einem erheblichen Maß an Rechtsunsicherheit konfrontiert, welchen auch durch das neue GNDG nur teilweise abgeholfen wird. In der Konsequenz wird oftmals stattdessen dann eine aufwändige und den wissenschaftlichen Mehrwert oft einschränkende rein einwilligungsbasierte Vorgehensweise gewählt, die in vielerlei Hinsicht den Forschungsnotwendigkeiten nicht hinreichend Rechnung tragen kann.

---

[28] § 46 Abs. 1 Nr. 2a LKHG BW: „soweit dies erforderlich ist … zur Durchführung medizinischer Forschungsvorhaben des Krankenhauses"; Art. 27 Abs. 4 S. 1 BayKRG: „soweit dies … zu Forschungszwecken im Krankenhaus oder im Forschungsinteresse des Krankenhauses erforderlich ist".

[29] Art. 27 Abs. 4 S. 2 BayKRG: Zu Forschungszwecken „können sie anderen Personen die Nutzung von Patientendaten gestatten, wenn dies zur Durchführung des Forschungsvorhabens erforderlich ist und die Patientendaten im Gewahrsam des Krankenhauses verbleiben."

[30] Im Einzelnen § 25 Abs. 1 LKG Berlin.

[31] § 31 BbgKHEG.

[32] S. etwa § 38 Abs. 1 LKHG SH, wonach Patientendaten für Forschungsdaten verarbeitet werden dürfen, „soweit die Patientin oder der Patient hinreichend aufgeklärt wurde und in die Datenverarbeitung für die Zwecke der wissenschaftlichen Forschung eingewilligt hat."

3. Rechtsunsicherheit hinsichtlich Anonymisierung

Das Problem der restriktiven und kaum rechtssicher handhabbaren Legitimationsgrundlagen für eine Forschungsdatenverarbeitung wäre letztlich vernachlässigenswert, wenn sich das Konzept einer deutschen MIMIC (ebenso wie in den USA) auf eine Anonymisierung der verarbeiteten Daten von Patientinnen und Patienten stützen könnte. Datenschutzrechtlich relevant ist eine Datenverarbeitung – auch im MIMIC-Kontext – stets nur insoweit, als es überhaupt zu einer Verarbeitung von *personenbezogenen* Daten kommt. Nach EG 26 der DS-GVO gelten die Grundsätze des Datenschutzes nicht für "anonyme Informationen", verstanden als „Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann." Explizit bezieht sich der Verordnungsgeber im Folgenden dann gerade auch auf eine Datenverarbeitung für Forschungszwecke. Soweit Daten im Rahmen einer MIMIC somit rechtssicher anonymisiert werden können, könnten sie auch als "Open Data" allgemein verfügbar gemacht werden.

Trotz der großen Bedeutung einer Unterscheidung zwischen personenbezogenen Daten einerseits und anonymisierten Daten finden sich in der DS-GVO allerdings keine greifbaren Regelungen, die eine klare und rechtssichere Differenzierung zwischen beiden Kategorien ermöglichen würden. Nach EG 26 der DS-GVO sollen für die Frage der Personenbeziehbarkeit „alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern." Ob wiederum Mittel „nach allgemeinem Ermessen wahrscheinlich zur Identifizierung der natürlichen Person genutzt werden", soll unter Heranziehung aller „objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand," beurteilt werden, „wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen sind."

In Ermangelung klarer gesetzlicher Vorgaben ist damit aber auch die Einordnung von Daten als anonymisiert mit einem erheblichen Maß an Rechtsunsicherheit behaftet – was dazu führt, dass in Deutschland der Weg über die Datenanonymisierung im Vergleich zu anderen Ländern nur selten beschritten wird, um eine Datennutzung für Forschungszwecke zu ermöglichen. Weiter verstärkt wird diese Zurückhaltung durch die immer noch sehr strikte Sichtweise der Datenschutzaufsichtsbehörden zum Personenbezug von Daten und die schon oben beklagte Tendenz hierzulande, generell bei Lösungsansätzen im Datenschutz einen überaus strengen und risikoaversen Beurteilungsmaßstab anzulegen.

Ganz anders präsentiert sich die Situation in den Niederlanden, obwohl hier mit der DS-GVO im Ausgangspunkt die gleichen rechtlichen Rahmenbedingungen gelten. Die Etablierung der niederländischen MIMIC (UMCdb) fußt zentral auf der Prämisse, dass die über die UMCdb zur Verfügung gestellten Daten rechtssicher anonymisiert worden sind und daher frei zur Verfügung gestellt werden können. Basis hierfür ist ein externes Audit, welches nach Durchführung eines „Reidentification Risk Assessment" zu dem Ergebnis gelangt ist, dass "unter Berücksichtigung aller Mittel, die nach vernünftigem Ermessen für eine Re-Identifizierung ver-

wendet werden können, eine Re-Identifizierung im Fall der UMCdb unwahrscheinlich ist, weshalb von anonymen Informationen im Kontext der DS-GVO ausgegangen werden kann".[33] Vorausgesetzt wird insoweit gerade nicht, dass Daten mit hundertprozentiger Sicherheit von einer De-Anonymisierung (Re-Identifizierung) geschützt sind. Es besteht vielmehr Einigkeit, dass sich eine solche 100-Prozent-Garantie eines Schutzes vor Re-Identifikation niemals herstellen lässt, dass eine solche Garantie aber auch keine Voraussetzung für die Annahme anonymisierter Daten ist.[33]

Im selben Sinne ist an sich auch hierzulande in der Diskussion anerkannt, dass eine sog. absolute („vollkommene") Anonymisierung häufig nicht möglich ist, es datenschutzrechtlich aber auch gar nicht erforderlich ist, dass eine Wiederherstellung des Personenbezugs von Daten für niemanden mehr möglich ist. Auch der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (BfDI) sieht es in seinem Positionspapier zur Anonymisierung regelmäßig als ausreichend an, dass "eine Re-Identifizierung praktisch nicht durchführbar ist, weil der Personenbezug nur mit einem unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft wiederhergestellt werden kann" (sog. faktische Anonymisierung).[34]

Im Ausgangspunkt gelten hierzulande damit also die gleichen Maßstäbe wie auch in den Niederlanden. Anders als in den Niederlanden werden daraus allerdings keine verwertbaren Schlussfolgerungen gezogen. Woran es fehlt, ist in einem zweiten Schritt die Klärung, unter welchen konkreten Voraussetzungen – im positiven Sinne – von einer rechtssicheren Datenanonymisierung ausgegangen und damit eine Forschungsdatenverarbeitung ermöglicht werden kann. Stattdessen verliert sich die weitere Diskussion wahlweise in allgemeinen Erörterungen zu den Schwierigkeiten einer verlässlichen Anonymisierung oder gar in grundsätzlichen Zweifeln an der Anonymisierung als tragfähiges Schutzkonzept.[35] Symptomatisch ist insoweit auch das Positionspapier des BfDI, nach dem eine "valide Anonymisierung – je nach Art der zu anonymisierenden Daten und Kontext der Verarbeitung – eine Herausforderung für den jeweiligen Verantwortlichen" sei und deshalb „nicht vorschnell von einer hinreichenden Anonymisierung ausgegangen werden" dürfe.[36] Wenn sich selbst eine Aufsichtsbehörde wie der BfDI in einem „Positionspapier zur Anonymisierung" letztlich nicht konkreter dazu äußert, ob und ggf. unter welchen Voraussetzungen eine rechtssichere Anonymisierung in Betracht

[33] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, Bailey H, Kesecioglu J, Cecconi M, Churpek M, Clermont G, van der Schaar M, Ercole A, Girbes ARJ, Elbers PWG; Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med. 2021 Jun 1;49(6):e563-e577.

[34] BfDI. Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche. 29.6.2020; S. 4 unter Verweis u.a. auf EuGH, Urt. v. 19.10.2016 – C-582/14 – Breyer.

[35] Zu letzterem Aspekt s. etwa Burkert C, Federrath H, Marx M, Schwarz M. Stellungnahme im öffentlichen Konsultationsverfahren des BfDI zum Thema „Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche". [Internet] 20.03.2020 [zitiert 20. Dezember 2023]. Verfügbar unter: https://svs.informatik.uni-hamburg.de/publications/2020/2020-03-20-Burkert-Stellungnahme-BfDI.pdf.

[36] BfDI. Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche. 29.06.2020; S. 4.

kommt, dürfte es bis auf Weiteres unrealistisch sein, dass sich Forschende auf die Unsicherheit einlassen, über den Weg der Anonymisierung für eine Verfügbarkeit von Forschungsdaten zu sorgen.

II. Bedarfsanalyse: Bedarf es einer deutschen MIMIC?

Der spezifische Ansatz einer MIMIC-Datenbank zeichnet sich dadurch aus, de-identifizierte Daten fachöffentlich zu machen. Alternativ stehen aber auch andere Modelle zur Verfügung, um sensible Daten der Forschung zugänglich zu machen, beispielsweise Datentreuhandstellen, föderiertes Lernen oder sichere Verarbeitungsumgebungen bei Datenzugangsstellen. Es stellt sich daher die Frage, ob eine vollständiger Reproduktion eines „Open Data"-Ansatzes nach dem Vorbild von MIMIC überhaupt einen Mehrwert hat oder der Aufbau einer deutschen MIMIC nicht besser einem auf Grundlage der MIMIC-Erfahrungen und der europäischen Rechtskultur modifizierten Ansatz folgen sollte.

1. Grundsätzliche Relevanz einer *deutschen* MIMIC?

Zur Frage der Notwendigkeit oder Sinnhaftigkeit einer "deutschen MIMIC" ist zunächst unabhängig von deren konkreter organisatorischer und technischer Ausprägung (insb. im Sinne eines "Open Data"-Ansatzes) festzustellen, dass die niederschwellige Ermöglichung der Sekundärnutzung von RWD für große, repräsentative Teile der deutschen Patientinnen- und Patientenversorgung dringend geboten erscheint, und zwar nicht nur für die Intensivmedizin. Jedenfalls ist es keine Option, sich hierzulande für die Forschung darauf zu beschränken, auf die Daten der MIMIC aus den USA zurückzugreifen. Dies begründet sich einerseits aus der Nicht-Vergleichbarkeit der Gesamtsituation der USA einerseits bezüglich der Patientinnen- und Patienteneigenschaften z.B. ethnischer Zugehörigkeit und Komorbiditäten, andererseits aber auch aus den z.T. abweichenden Rahmenbedingungen und Vorgehensmodellen in der Gesundheitsversorgung selbst.

Hinzu kommen die intrinsischen Einschränkungen, mit denen MIMIC in seiner aktuellen Ausprägung behaftet ist: Auch für die Patientinnen- und Patientenpopulation und Versorgungspraxis der USA ist MIMIC notwendigerweise nur begrenzt repräsentativ, da die Datenquelle im Wesentlichen monozentrisch ist und somit weder bzgl. regionaler Variabilität innerhalb des Landes noch bzgl. interinstitutioneller Variabilität informativ sein kann.

2. Repräsentativitäts- und Diversitätsanforderungen sowie Implikationen für die Planung und Steuerung der Entwicklung des Gesundheitssystems

Um für in Deutschland behandelte Patientinnen und Patienten maximal relevante Forschung und Qualitätssicherung organisieren zu können, sind Datenbestände erforderlich, die die Eigenschaften und die dynamische Entwicklung der deutschen Patientinnen- und Patientenpopulation und der deutschen Versorgungspraxis adäquat repräsentieren. Idealerweise wird dies erreicht, indem Datenquellen aus allen Regionen und allen Versorgungsstufen in repräsentativer Gewichtung bzw. Verteilung in die Datenbasis einfließen.[37] So könnte nicht nur

[37] Auf die aus dieser Anforderung resultierenden Herausforderungen gehen wir unter C.III.4 näher ein.

mehr translational für die deutsche Bevölkerung relevante Forschung ermöglicht werden, sondern auch eine qualitätssichernde und steuernde Ausprägung eines "lernenden Gesundheitssystems" gestaltet werden. Auf einer repräsentativen Datengrundlage lassen sich nämlich prädiktive Modelle kalibrieren, die einer risikoadjustierten Ergebnisqualitätsmessung anhand patientinnen- und patientenzentrierter Ergebnisindikatoren den Weg bahnen und damit eine essentielle Voraussetzung für die praxisgerechte und patientinnen- und patientenorientierte Ausprägung einer leistungsbasierten Umgestaltung der erlösbasierten Anreizsysteme im deutschen Gesundheitssystem schaffen könnte. Dies könnte zur dringend gebotenen Reduktion der administrativen Aufwände für die Durchführung und Kontrolle der Abrechnung der Leistungserbringer beitragen, deren Exazerbation über die letzten Jahrzehnte z.B. in Form von jährlich aktualisierten Abrechnungsregelwerken im für den stationären Bereich relevanten System der diagnosebezogene Fallgruppen mit mehr als 5000 Seiten massive rein erlössicherungsbedingte Dokumentationsaufwände induziert. Diese Entwicklungen resultieren u.a. aus der ständigen Notwendigkeit der kleinteiligen Nachjustierung der zentral definierten Anreizsysteme aufgrund festgestellter, bei einem solchen zentralisierten Ansatz fast unvermeidlicher Konflikte zwischen Erlösanreizen und medizinischen und für die Patienteninnen und Patienten sinnvollen Entscheidungen in den konkreten, hochvariablen Behandlungskonstellationen.

Sollte eine für die deutsche intensivmedizinische Versorgungslandschaft repräsentative deutsche MIMIC nicht realisierbar sein, könnte durch Umsetzung einer reduzierten Ausprägung, z.B. durch einen fokussierten Aufbau im Rahmen einer Kooperation zwischen wenigen großen Universitätskliniken mit geeigneten infrastrukturellen Voraussetzungen, ein MIMIC-analoger Datensatz geschaffen werden, der die Patientinnen- und Patientenpopulation und Versorgungssituation in Deutschlands zumindest etwas besser approximiert als dies für die rein monozentrische, ursprüngliche MIMIC der Fall ist. Eine Anwendung im Sinne des oben beschriebenen potentiell nachhaltigen und gesamtsystemisch wirksamen Innovationsökosystems oder gar eine Nutzung zur Verschlankung und Ergebnisfokussierung der intensivmedizinischen Abrechnungssystematik käme in diesem Fall aber aufgrund der fehlenden Repräsentativität bzgl. Patientinnen- und Patientenkollektiv und der fehlenden Möglichkeit einer Ausweitung in die Breite aus offensichtlichen Gründen nicht in Frage. Dies bedeutet wiederum, dass auch keine indirekte Refinanzierung über Qualitätseffekte und die Vereinfachung administrativer Prozesse möglich wäre. Selbst für explorative wissenschaftliche Analysen erscheint die Wirtschaftlichkeit einer solchen Investition in einem monozentrischen Setting sehr fraglich, da solche explorativen Analysen letztlich auch auf international bereits verfügbaren Datensätzen durchgeführt werden könnten und für die Überprüfung der Generalisierbarkeit auf die in Deutschland relevante Versorgungssituation unverändert geeignete Werkzeuge fehlen würden.

Eine funktionierende Erlösincentivierung auf Grundlage repräsentativer RWD sauber risikoadjustierter, patientinnen- und patientenzentrierter Ergebnisqualität könnte perspektivisch die aus medizinischer Sinnhaftigkeit, Patientinnen- und Patienteninteressen und wirtschaftlichen Erwägungen resultierenden Entscheidungstendenzen potentiell systematisch, dezentral und effizient, da bzgl. des Einzelvorganges mit geringerem Dokumentations- und Kontrollaufwand verbunden, in Übereinstimmung bringen. Ein Beispiel für eine solche datengetrieben risikoad-

justierte Qualitätsmessung in der Intensivmedizin, dort allerdings aktuell basierend auf registerartig dediziert erfassten strukturierten Behandlungsdaten, ist z.B. das NICE-System in den Niederlanden[38].

Funktionierende RWD-Nutzungsstrukturen bilden aber gerade in der deutschen, sektoral und organisatorisch zergliederten Versorgungslandschaft einen unverzichtbaren Schritt auf dem Weg zur Realisierung eines sinnvollen qualitätsorientierten Vergütungsverfahrens, da hier vor allem auch das Problem der Attributierung des patientinnen- und patientenzentrierten Behandlungsergebnisses auf die am konkreten Behandlungsvorgang sequentiell beteiligten intersektoralen Leistungserbringer gelöst werden muss: Wie viel hat etwa in einem konkreten Behandlungsfall der primär diagnostizierende und nachbehandelnde Hausarzt zur z.B. 6 Monate nach Behandlungsabschluss erhobenen Lebensqualität der Patientin bzw. des Patienten beigetragen, wie viel der in die Primärdiagnostik einbezogene niedergelassene Facharzt, wie viel das die primäre Operation verantwortende Krankenhaus der Regelversorgung, wie viel das wegen aufgetretener Komplikationen einbezogene Universitätsklinikum, und wieviel schließlich die Reha-Einrichtung? Große Datenbasen von RWD erscheinen aktuell als einzig plausibler Ansatz, eine solche Attributierung überhaupt systematisch zu versuchen.

3. Möglichkeit und Mehrwert eines vollständigen “Open Data”-Ansatzes

Der mögliche Mehrwert eines “Open Data”-Ansatzes gegenüber z.B. auf Datentreuhandstellen, föderiertem Lernen oder sicheren Verarbeitungsumgebungen bei Datenzugangsstellen basierenden, alternativen Ansätzen zur Nutzbarmachung von RWD (siehe nächster Abschnitt) hängt primär davon ab, welchen Einschränkungen die alternativen Vorgehensmodelle unterliegen. Oder umgekehrt formuliert: Je praktikabler und niederschwellig und flexibel nutzbarer diese Alternativen ausgestaltet werden können, desto geringer ist der Mehrwert eines Open Data-Ansatzes in Deutschland.

Zunächst einmal ist davon auszugehen, dass wohl praktisch jede restriktivere Lösung gegenüber einem Open Data-Ansatz, bei dem der Datennutzer völlig frei entscheiden kann, wo und wie er die Daten verarbeitet, mindestens gewisse Effizienzverluste bedingen wird. Bei großen Datenmengen wie z.B. Ganzgenomsequenzen oder Bilddaten aus der Pathologie kann dieser vermeintliche Nachteil allerdings dadurch relativiert oder gar ins Gegenteil verkehrt werden, dass die Verarbeitung solcher Daten oft ohnehin nur auf spezialisierter Rechnerinfrastruktur sinnvoll möglich ist. Wenn diese z.B. in einem Trusted Research Environment bereitgestellt würde, entfallen Transfer und redundante Speicherung der großen Datenmengen und der Effizienzverlust verkehrt sich u.U. sogar ins Gegenteil. Zu berücksichtigen ist auch, dass ein Teil dieser Datenarten besonders schützenswert ist. Genetische Daten etwa tragen nicht nur in ihrem tatsächlichen Informationsgehalt bisher nur in kleinen Teilen verstandene Informationen über den betreffenden Menschen, sondern sind sicherlich mit den richtigen Analyseverfahren auch informativ bzgl. z.B. Leistungspotential, Vulnerabilität und anderen sensiblen Eigen-

[38] van de Klundert N, Holman R, Dongelmans DA, de Keizer NF. Data Resource Profile: The Dutch National Intensive Care Evaluation (NICE) Registry of Admissions to Adult Intensive Care Units. International Journal of Epidemiology. 1. Dezember 2015;44(6):1850–1850h.

schaften - und dies nicht nur für die betroffene Person selbst, sondern auch für deren Verwandte. Solche besonders schützenswerten Daten nicht ohne Not unkontrolliert zu replizieren und zu streuen erscheint nur vernünftig.

Zusammenfassend besteht jedenfalls ein großer Bedarf, ein deutsches MIMIC-Analog zu schaffen, welches die deutsche Bevölkerung und die Verhältnisse im deutschen Gesundheitssystem adäquat abbildet. Hierdurch würde nicht nur für die deutsche Bevölkerung maximal relevante intensivmedizinische Forschung ermöglicht, sondern auch ein wichtiger Beitrag zum Aufbau des lernenden Gesundheitssystems der Zukunft geleistet. Wichtig ist hierbei vor allem die niederschwellige Nutzbarkeit der Daten und die Zukunftssicherheit des Konzeptes auch mit Blick auf die immer wichtiger werdenden hochdimensionalen Datenquellen. Letztere Datenquellen mit ihrem besonderen Risikopotential sind zugleich dann aber auch ausschlaggebend dafür, dass ein MIMIC-Analog basierend auf einem *vollständigen* Open Data-Ansatz in Deutschland möglicherweise nicht die optimale Lösung darstellt.

III. Umsetzung: Wie ließe sich eine „deutsche MIMIC“ bauen?

Im Folgenden wird die Frage nach Optionen zur Umsetzung einer “deutschen MIMIC” anhand verschiedener Achsen detailliert bearbeitet. Zuerst werden grundsätzliche Umsetzungsszenarien skizziert, mit einem besonderen Fokus auf Anonymisierungsverfahren auf Datenebene, da diese einen wesentlichen Baustein der MIMIC darstellen und als ein Kerncharakteristikum angesehen werden können. Die identifizierten Szenarien werden bzgl. Chancen und Risiken analysiert. Anschließend werden notwendige rechtliche Rahmenbedingungen für die Umsetzungsszenarien behandelt, Einflussmöglichkeiten der Bundesbehörden in Richtung einer Entwicklung entsprechender Strukturen beschrieben und Ressourcenbedarfe abgeschätzt.

1. Analyse grundsätzlicher Umsetzungsszenarien zum Schutz der Privatsphäre

Sensible medizinische Daten können im Rahmen der Offenlegung für Forschungszwecke sowohl auf Datenebene ("Anonymisierung") als auch auf Prozessebene geschützt werden. Ansätze auf Datenebene wurden unter anderem für die Anonymisierung tabellarischer, physiologischer, Text- und Bilddaten vorgeschlagen. Auf Prozessebene können Verfahren wie Föderation oder Trusted Research Environments eingesetzt werden. Beide Ansätze stehen jedoch in der Praxis großen Herausforderungen gegenüber, da sie die Nützlichkeit der Daten oder der umsetzbaren Forschungsprozesse deutlich beeinflussen können. In der Praxis ist es deshalb meist sinnvoll, Methoden auf der Daten- und Prozessebene zu kombinieren, was aber wiederum zu Herausforderungen bei der Passgenauigkeit mit rechtlichen Anforderungen führt (siehe Abschnitte C.IV und C.V). Neueste Entwicklungen insbesondere aus den Bereichen Datensynthetisierung und Differential Privacy können neue Möglichkeiten eröffnen, sind jedoch ebenfalls mit spezifischen Herausforderungen und Unsicherheiten behaftet.

a) Szenario 1: Fachöffentliche Bereitstellung durch Anonymisierung auf Datenebene

Bei Szenario 1 liegt der Fokus auf der Anonymisierung und Synthetisierung auf Datenebene. Szenario 1 kommt dem Design der amerikanischen MIMIC am nächsten. Auf dieser Grundlage könnte in weitestmöglichem Umfang die "Open Data”-Idee von MIMIC umgesetzt werden.

Aufgezeigt werden soll im Folgenden, welche Vielzahl an Instrumentarien in der Praxis mittlerweile bereitsteht, um im Rahmen eines MIMIC-Systems die Risiken einer Re-Identifizierung von Daten so gering wie möglich zu halten und einen robusten Grundschutz zu gewährleisten. Daten werden hierbei so stark modifiziert oder synthetisiert, dass ein Personenbezug nur noch sehr schwer herstellbar ist. Die Auswirkungen dieser Veränderungen auf die wissenschaftliche Nutzbarkeit werden allerdings stark von konkreten Nutzungsszenario abhängen und nicht immer vernachlässigbar sein. Um eventuelle Restrisiken, die trotz der Anonymisierung bestehen könnten, zu minimieren, werden die Daten ausschließlich an vertrauenswürdige Parteien herausgegeben. Der Zugang zu den Daten erfolgt auf Antragsbasis und ist an die Unterzeichnung einer Nutzungsvereinbarung gebunden, die die Einhaltung bestimmter Sicherheits- und Datenschutzstandards vorschreibt.

Welche konkreten Anforderungen an die Anonymisierungsverfahren zu stellen sind, hängt im Ausgangspunkt von der jeweiligen Beschaffenheit der verschiedenen Datenarten ab. Diesbezüglich können die wesentlichen Datenarten in MIMIC in folgende Kategorien zusammengefasst werden:

- **Tabellarische Daten** (u.a. Demographie und klinische Parameter)
- **Physiologische Daten** (u.a. Biosignale und Zeitreihen)
- **Texte** (u.a. Notizen, Arztbriefe, Radiologiebefunde)
- **Bilddaten** (nicht direkt enthalten, nur per Zusatzantrag verfügbar)

Bei tabellarischen Daten, die demografische Informationen und klinische Parameter umfassen, liegt der Fokus auf der Verallgemeinerung oder Entfernung direkter und indirekter Identifikatoren wie Namen oder exakter Adressen[39]. Physiologische Daten, insbesondere Zeitreihen, bergen das Risiko der Identifizierung durch einzigartige Muster[40]. Die Anonymisierung muss hier darauf abzielen, diese Muster zu verschleiern, ohne die medizinische Relevanz der Daten zu beeinträchtigen. Textdaten wie Notizen und Berichte enthalten oft versteckte persönliche Informationen in freiem Text. Hier sind Techniken des Natural Language Processing notwendig[41]. Bilddaten schließlich enthalten oft visuell identifizierbare Merkmale. Die Anonymisierung muss hier sowohl sichtbare Merkmale als auch potenziell sensible Metadaten berücksichtigen[42].

aa) Entfernung direkt identifizierender Merkmale

Allen Verfahren geht voraus, dass direkt identifizierende Merkmale entfernt werden müssen und gleichzeitig der Bezug zwischen unterschiedlichen Daten und Datenarten beibehalten

[39] El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015 Mar 20;350:h1139. doi: 10.1136/bmj.h1139. PMID: 25794882; PMCID: PMC4707567.
[40] Jafarlou S, Rahmani AM, Dutt N, Mousavi SR. ECG Biosignal Deidentification Using Conditional Generative Adversarial Networks. Annu Int Conf IEEE Eng Med Biol Soc. 2022 Jul;2022:1366-1370. doi: 10.1109/EMBC48229.2022.9872015. PMID: 36086579.
[41] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):550-63.
[42] Muschelli J. Recommendations for Processing Head CT Data. Front Neuroinform. 2019 Sep 4;13:61. doi: 10.3389/fninf.2019.00061. PMID: 31551745; PMCID: PMC6738271.

werden muss[43]. Dazu gehören spezifische Informationen wie Namen, Adressen, Telefonnummern, Versicherungsnummern und andere eindeutige Identifikatoren, wie Personalnummern. Um jedoch den Zusammenhang zwischen verschiedenen Datensätzen und Datenarten aufrechtzuerhalten, ohne die Identität der Patientinnen und Patienten preiszugeben, ist die Einführung von Surrogatidentifikatoren notwendig. Diese künstlichen Kennungen erlauben es, wenn sie konsistent eingesetzt werden, Daten aus verschiedenen Quellen zu einer Person zuzuordnen, ohne dass Rückschlüsse auf die reale Identität möglich sind.

bb) Umgang mit Zeitangaben

Der Umgang mit Zeitstempeln in medizinischen Daten ist eine besondere Herausforderung, da sie Muster enthalten können, die zur Re-Identifizierung von Patientinnen und Patienten genutzt werden könnten. Moderne Verfahren zur Anonymisierung von Zeitangaben zielen darauf ab, diese Informationen zu verschleiern, ohne ihren Nutzen wesentlich zu beeinträchtigen. Eine gängige Methode ist das Verschieben von Zeitstempeln, wobei alle relevanten Datenpunkte um einen konsistenten Zeitraum verschoben werden[44]. Dies könnte bedeuten, dass alle Zeitstempel einer Patientin bzw. eines Patienten um dieselbe Anzahl von Tagen, Wochen oder Monaten in die Vergangenheit oder Zukunft verlagert werden. Eine weitere Methode ist die Verwendung relativer Zeitangaben, bei der das Datum und die Uhrzeit durch die Dauer seit einem bestimmten Ereignis (z.B. Aufnahmedatum) ersetzt werden. Wichtig ist dabei, dass Änderungen an Zeitangaben über unterschiedliche Datensätze zur selben Person hinweg konsistent sein müssen. Bei jedem Verfahren bestehen darüber Herausforderungen bzgl. des Erhalts forschungsrelevanter Informationen, da Veränderungen an zeitlichen Daten bspw. Informationen zur Saisonalität oder zur Unterscheidung von Wochen und Werktagen beeinflussen können.

cc) Anonymisierung tabellarischer Daten

Das bekannteste Risikomodell für tabellarische Daten ist die k-Anonymität, die das Re-Identifizierungsrisiko über die Unterscheidbarkeit von anderen Individuen bestimmt und eine Gruppengröße von „k“ sicherstellt[45]. Eine Erweiterung der k-Anonymität ist die l-Diversity, die sicherstellt, dass sensible Merkmale innerhalb der Gruppen von ununterscheidbaren Individuen möglichst divers sind, um Attribute Disclosure, also die Ableitung sensibler Informationen über Personen auch ohne direkte Zuordnung (siehe Abschnitt C.IV) zu verhindern[46]. Ein weiteres Modell, die t-Closeness, verlangt, dass die Verteilung der sensiblen Merkmale in jeder Gruppe der Gesamtverteilung im Datensatz ähnlich sein muss, um subtilere Formen der Attribute

[43] Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, Shirey-Rice J, Kirby J, Harris PA. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014 Dec;52:28-35.
[44] Hripcsak G, Mirhaji P, Low AF, Malin BA. Preserving temporal relations in clinical data while maintaining privacy. J Am Med Inform Assoc. 2016 Nov;23(6):1040-1045.
[45] Sweeney L. K-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst. 2002;10(5):557-70.
[46] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data (TKDD). 2007;1(1):3.

Disclosure zu verhindern[47]. Darüber hinaus wurden viele weitere Modelle und Verfahren vorgeschlagen[48].

Grundlegende Verfahren zur Veränderung von Daten zum Schutz der Privatheit umfassen die Löschung und Generalisierung von Einzelwerten, Variablen oder Datensätzen[49]. Die Aggregation, bei der Daten zusammengefasst werden, trägt ebenfalls zur Reduzierung der Spezifität der Information und damit des Re-Identifizierungsrisikos bei[50]. Die Stichprobenziehung, also die Verwendung eines zufälligen Teils der Gesamtdaten, kann ebenfalls das Risiko senken[51]. Eine weitere bekannte Methode ist das sogenannte „Swapping", bei dem bestimmte Werte mit einer gewissen Wahrscheinlichkeit, die wiederum abhängig von der Häufigkeit der Werte ist, zwischen Datensätzen ausgetauscht werden[52]. Darüber hinaus wurden viele weitere Verfahren vorgeschlagen[53].

dd) Anonymisierung von Bilddaten

Grundlegende Methoden für die Anonymisierung von Bilddaten umfassen die Löschung, Verpixelung oder Verunschärfung von identifizierbaren Merkmalen wie Gesichtern im zwei- oder dreidimensionalen Raum[54]. Darüber hinaus können Geometrieänderungen angewendet werden, wie das Zuschneiden oder Drehen von Bildern, um eine eindeutige Zuordnung zu verhindern. Sensible Bereiche in Bildern, die vertrauliche Informationen enthalten könnten, können durch Ausblenden oder Maskieren geschützt werden. Zusätzlich zu diesen Methoden werden auch Techniken wie das Hinzufügen von Rauschen oder die Veränderung der Bildauflösung verwendet, um eine Re-Identifizierung zu erschweren.

ee) Anonymisierung von Textdaten

Grundlegende Techniken zur Anonymisierung von Textdaten umfassen die Erkennung und Entfernung identifizierender Textbestandteile, sogenannter Tokens. Diese sowie spezifische medizinische Begriffe oder sensitive Informationen, die zur Identifizierung führen könnten,

---

[47] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007. IEEE Computer Society.

[48] Wagner I, Eckhoff D. Technical Privacy Metrics: A Systematic Survey. ACM Comput Surv. 2018;51(57):1-57:38.

[49] Sweeney L. K-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst. 2002;10(5):557-70.

[50] Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. Data Knowl Eng. 2005;55(3):301-22.

[51] Willenborg L, de Waal T. Statistical disclosure control in practice. Springer Science & Business Media; 2001.

[52] Willenborg L, de Waal T. Statistical disclosure control in practice. Springer Science & Business Media; 2001.

[53] Templ M. Statistical disclosure control for microdata. Cham: Springer; 2017.

[54] Clunie DA, Flanders A, Taylor A, Erickson B, Bialecki B, Brundage D, Gutman D, Prior F, Seibert JA, Perry J, Gichoya JW, Kirby J, Andriole K, Geneslaw L, Moore S, Fitzgerald TJ, Tellis W, Xiao Y, Farahani K, Luo J, Rosenthal A, Kandarpa K, Rosen R, Goetz K, Babcock D, Xu B, Hsiao J. Report of the Medical Image De-Identification (MIDI) Task Group - Best Practices and Recommendations. ArXiv [Preprint]. 2023 Apr 1:arXiv:2303.10473v2.

werden idealerweise nicht einfach gelöscht, sondern durch konsistente zufällig generierte Informationen ersetzt, um den Schutz weiter zu erhöhen[55].

In den vergangenen Jahren haben automatisierte Textverarbeitungsverfahren, insbesondere solche, die auf Computerlinguistik, auch "Natural Language Processing" (NLP), basieren, stark an Bedeutung gewonnen. NLP-Modelle können dazu verwendet werden, große Mengen von medizinischen Texten effizient zu analysieren und personenbezogene Informationen zu identifizieren. Diese automatisierten Ansätze erleichtern den Anonymisierungsprozess erheblich und reduzieren den manuellen Aufwand. Es können folgende wesentlichen Arten von Methoden unterschieden werden:

- **Regelbasiertes NLP**: Hierbei werden vordefinierte Regeln und Muster verwendet, um personenbezogene Informationen zu identifizieren und zu maskieren.

- **Machine-Learning-basiertes NLP**: Diese Ansätze nutzen maschinelles Lernen, um Tokens in den Texten zu erkennen und personenbezogene Informationen automatisch zu entfernen.

- **Deep Learning-basiertes NLP**: Diese Variante von Machine-Learning-basierten Verfahren nutzen Tiefe neuronale Netzwerke, einschließlich rekurrenter neuronaler Netzwerke (RNNs) und Transformer-Modelle, um komplexe Abhängigkeiten im Text zu erfassen und so personenbezogene Informationen noch treffsicherer zu entfernen.

- **Transfer Learning-basiertes NLP**: Diese Variante von Deep-Learning-basierten Verfahren nutzen vortrainierte NLP-Modelle, die auf großen Textdatenmengen generiert worden sind, und führen Feinabstimmungen in Bezug auf die Aufgabe der Anonymisierung von Texten durch.

Trotz der Fortschritte in der automatisierten Anonymisierung ist im Regelfall eine sorgfältige Prüfung der anonymisierten Texte unerlässlich.

ff) Differential Privacy

Differential Privacy ist eine moderne Methode der Anonymisierung bzw. Verarbeitung von Daten mit Privatheitsgarantien[56,57]. Die Technik basiert darauf, randomisierte Funktionen auf Datensätze anzuwenden, wobei jeder Datensatzeintrag die Informationen genau einer Person enthalten muss. Da die Funktionen randomisiert sind, sind die möglichen Ausgaben eine Wahrscheinlichkeitsverteilung. Eine gängige Methode zur Randomisierung von Funktionen, um die Differential Privacy-Eigenschaft herzustellen, ist das hinzunehmen einer Verrauschungsfunktion[58].

---

[55] Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, Hirschman L. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):342-8. doi: 10.1136/amiajnl-2012-001034. Epub 2012 Jul 6.
[56] Dwork C. Differential Privacy. In: 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006).
[57] Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data Analysis. In: Proceedings of the Third Theory of Cryptography Conference (TCC 2006).
[58] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Found Trends Theor Comput Sci. 2014;9(3–4):211-407.

Die Kernidee von Differential Privacy besteht darin, dass die Wahrscheinlichkeitsverteilung der Ausgaben für zwei beliebige, aber fast identische Datensätze – die sich nur in einem Eintrag unterscheiden – nur minimal variieren darf. Das Ausmaß dieser Variation wird durch den Parameter "Epsilon", das sogenannte "Privacy Budget", bestimmt[56]. Diese Eigenschaft stellt sicher, dass grundsätzliche Erkenntnisse über den gesamten Datensatz gewonnen werden können, ohne dass die Daten einzelner Personen einen direkten Einfluss darauf haben. Dadurch wird verhindert, dass spezifische Informationen über Einzelpersonen preisgegeben werden.

Obwohl Differential Privacy von vielen Methodikern als Goldstandard für den Schutz der Privatheit bei der Verarbeitung personenbezogener Daten angesehen wird, bestehen in der Praxis große Herausforderungen. Das Konzept funktioniert relativ gut, wenn die Funktionen Daten stark aggregieren, also beispielsweise bei der Zählung von Patientinnen und Patienten. Herausforderungen bestehen hingegen, wenn die Funktionen detaillierte Ergebnisse produzieren. Hintergrund ist, dass die "Sensibilität" einer Funktion, also das Ausmaß, in dem die Daten einer einzelnen Person das Ergebnis beeinflussen können, die Intensität des erforderlichen Rauschens bestimmt. Bei Funktionen, die Ergebnisse mit einer hohen Informationsdichte produzieren, also beispielsweise Kohortencharakterisierungen statt einfachen Fallzahlen, ist die Sensibilität im Regelfall sehr hoch, was wiederum ein starkes Rauschen erfordert und die Nützlichkeit des Ergebnisses stark einschränkt.

Um diesen Herausforderungen zu begegnen, wurden viele schwächere Varianten von Differential Privacy oder nicht-strikte Implementierungen vorgeschlagen[59]. Diese bieten aber meist keine starken Schutzgarantien, was der Kernidee zuwiderläuft. Zudem erfordert der Einsatz von randomisierten Funktionen bei der Datenanalyse spezielle statistische Verfahren und Schulungen, da die Ergebnisse eben nicht exakt, sondern verrauscht sind. Aus diesen Gründen wird Differential Privacy beim Schutz von Daten in der medizinischen Forschung bisher kaum genutzt. Die Methode findet aber im Bereich der künstlichen Intelligenz (KI) und der Datensynthetisierung zunehmend Einsatz[60], allerdings auch hier mit großen Einschränkungen.

gg) Synthetische Daten

Die Datensynthetisierung ist ein Anonymisierungsverfahren, bei dem Schutz nicht durch Modifikation bestehender Daten, sondern durch Generierung neuer, künstlicher Daten erzeugt wird. Traditionelle Methoden der Synthetisierung setzen auf explizite statistische Modellierung der Eigenschaften des Quelldatensatzes[61]. Moderne Verfahren nutzen hingegen KI bzw. maschinelles Lernen, um wesentliche Eigenschaften der Originaldaten zu erlernen und dann zur Generierung neuer Daten zu nutzen[62]. Da es sich bei synthetischen Daten um künstliche Daten handelt, bestehen grundsätzlich große Vorteile bzgl. der Privatheit. So sind gut syntheti-

[59] Desfontaines D, Pejó B. SoK: Differential Privacies. arXiv preprint arXiv:1906.01337. 2019.
[60] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep Learning with Differential Privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.
[61] Nowok B, Raab GM, Dibben C. Synthpop: Bespoke Creation of Synthetic Data in R. J Stat Softw. 2016;74(11):1–26.
[62] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS 2014).

sierte Daten im Regelfall nicht mit einem Risiko von Identity Disclosure, also der direkten Zuordnung zu einer Person, behaftet. Allerdings besteht weiterhin die Gefahr von Attribute Disclosure oder Membership Disclosure (siehe Abschnitt C.IV)[63].

Die Synthetisierung von medizinischen Texten ist ein anspruchsvoller Prozess, der sich bisher nicht breit durchgesetzt hat. Wesentliche Herausforderungen liegen in der korrekten Wiedergabe medizinischer Fakten, sowie der Einhaltung der spezifischen sprachlichen Nuancen, die in der medizinischen Kommunikation verwendet werden. Forschungen auf diesem Gebiet haben gezeigt, dass der Einsatz von Deep-Learning-Modellen, insbesondere solchen, die auf RNNs und Transformern basieren, vielversprechend ist[64].

Die Synthetisierung von medizinischen Bilddaten, wie Röntgenbildern oder Magnetresonanztomographie (MRT)-Scans kann mittels Deep-Learning-Modellen umgesetzt werden. Ein spezifisches Verfahren, das häufig zur Anwendung kommt, sind Generative Adversarial Networks (GANs). GANs bestehen aus zwei Netzwerken, einem Generator und einem Diskriminator, die gegeneinander antreten, um die Qualität der generierten Daten kontinuierlich zu verbessern[62]. Ein Vorteil der Verwendung von Deep Neural Networks in der medizinischen Bildsynthese liegt in ihrer Fähigkeit, hochdetaillierte und realistische Bilder zu produzieren, die schwer von echten Aufnahmen zu unterscheiden sind[65]. Ein Nachteil dieser Methoden ist jedoch, dass sie eine große Menge an Trainingsdaten benötigen und manchmal Anomalien oder Artefakte erzeugen können, die in realen medizinischen Bildern nicht vorhanden sind[66].

Auch für die synthetische Generierung von Biosignaldaten wurden entsprechende Methoden vorgeschlagen[67]. Hier sind neben modernen Transformer-Modellen auch bereits länger bekannte Long Short-Term Memory (LSTM)-Netzwerk relevant[68]. Beide Verfahren sind besonders geeignet, um zeitabhängige Merkmale zu erfassen. Es kann nicht nur erforderlich sein, die grundlegenden biologischen Prozesse zu modellieren, sondern es müssen potenziell auch die Variabilität und Rauschen, die in echten Biosignalen vorhanden sind, berücksichtigt werden[69].

Eine übergreifende Herausforderung bei der Synthetisierung verschiedener medizinischer Datentypen liegt in der Wahrung der Konsistenz. Aktuelle Ansätze sind beispielsweise nur beschränkt in der Lage, zeitliche Veränderungen und Entwicklungen im Gesundheitszustand von Patientinnen und Patienten zu berücksichtigen und zeitliche Muster sowie Abhängigkeiten

[63] Stadler T, Oprisanu B, Troncoso C. Synthetic data–anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22); 2022. p. 1451-1468.
[64] Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, et al. Multi-layer representation learning for medical concepts. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[65] Nie D, Trullo R, Lian J, Wang L, Petitjean C, Ruan S, et al. Medical image synthesis with context-aware generative adversarial networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2017).
[66] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, Mukhopadhyay A. GANs for medical image analysis. Artif Intell Med. 2020;109:101938.
[67] Zhu FY, Ye F, Fu Y, Liu Q, Shen B. Electrocardiogram Generation with a Bidirectional LSTM-CNN Generative Adversarial Network. Sci Rep. 2019;9(1):6734.
[68] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng. 2019;16(5):051001.
[69] McSharry PE, Clifford GD, Tarassenko L, Smith LA. A dynamical model for generating synthetic electrocardiogram signals. IEEE Trans Biomed Eng. 2003;50(3):289-294.

beizubehalten. Darüber hinaus können nur eingeschränkt konsistente multi-modale Daten, bei denen beispielsweise Texte, Bilder und Biosignale, wechselseitig stimmig sind, erzeugt werden.

Wie bereits erwähnt, schützen synthetische Daten nicht unbedingt vor Attribute Disclosure oder Membership Disclosure. Eine häufig eingesetzte Methode zum weiteren Schutz der Privatheit bei synthetischen Daten ist Differential Privacy im Rahmen des Trainingsprozesses. Die Anforderungen von Differential Privacy vertragen sich grundsätzlich gut mit dem Einsatz beim Training von KI- und Machine Learning-Modellen, da es sowieso wünschenswert ist „Memorization" und „Overfitting" zu verhindern, um die Generalisierbarkeit der Modelle zu verbessern. Allerdings ist die Sensibilität der Ausgaben der Trainingsfunktionen oft zu hoch und das notwendige Rauschen wäre deshalb zu stark. Aus diesem Grund wird häufig Rauschen ohne strikte Garantien eingesetzt und somit kein garantierter Schutz erreicht. Dies macht wiederum empirische "Privacy Tests" erforderlich, um den erzielten Schutz zu evaluieren.

hh) Privacy Tests

Für die Bewertung von Restrisiken bei anonymisierten oder synthetischen Daten sowie auch KI-Modellen im Generellen wurden eine ganze Reihe von Ansätzen für "Privacy Tests" vorgeschlagen. Ein wichtiges Beispiel ist der Ansatz der sogenannten Shadow Models, bei dem KI-Modelle unter Einsatz frei verfügbarer Informationen darauf hintrainiert werden, Datensätze mit oder ohne Informationen zu einer spezifischen Person zu unterscheiden, was dann wiederum genutzt werden kann, um aus zugänglichen anonymen oder synthetischen Daten abzuleiten, ob eine spezifische Person Teil des ursprünglichen Datensatzes war[70]. Dieser Ansatz wurde beispielsweise von Stadler und Kollegen zur Restrisikomessung bei anonymisierten und synthetisierten Daten eingesetzt[71] und von Murakonda und Shokri auf ein breites Spektrum von KI-Modellen ausgeweitet[72]. Eine weitere häufig eingesetzte Methode ist die sogenannte "Holdout-Analyse", bei der Teile der Eingabedaten zurückgehalten werden und anschließend der Unterschied der Ähnlichkeit der synthetischen oder anonymisierten Daten mit den Eingabedaten und zurückgehaltenen Daten analysiert und verglichen wird[73]. Giomi und Kollegen haben einen ähnlichen Ansatz vorgeschlagen, um Daten auf Restrisiken in Bezug auf "Aussondern", "Verknüpfung" und "Inferenz" (siehe Abschnitt C.IV) zu untersuchen[74].

ii) Abschließende Bewertung

Klassische Anonymisierungsverfahren stehen im Zentrum einer kontroversen Diskussion[75], insbesondere im Hinblick auf hochdimensionale und multimodale Daten. Die Schwierigkeit,

[70] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017 May; pp. 3-18. IEEE.
[71] Stadler T, Oprisanu B, Troncoso C. Synthetic data–anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22); 2022. p. 1451-1468.
[72] Murakonda SK, Shokri R. ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. arXiv preprint arXiv:2007.09339. 2020.
[73] Platzer M, Reutterer T. Holdout-based empirical assessment of mixed-type synthetic data. Front Big Data. 2021;4:679939.
[74] Giomi M, Boenisch F, Wehmeyer C, Tasnádi B. A unified framework for quantifying privacy risk in synthetic data. Proc Privacy Enhanc Technol. 2023;312–328.
[75] Narayanan A, Felten EW. No silver bullet: De-identification still doesn't work. White Paper. 2014;8.

mit herkömmlichen Methoden einen starken und nachweisbaren Schutz gegen Re-Identifizierung zu gewährleisten, bleibt eine zentrale Herausforderung. Gleichzeitig gibt es Evidenz, dass die Risiken einer Re-Identifizierung in der Praxis oft gering sind[76]. Verfahren wie Maskierung, Generalisierung und Pseudonymisierung stellen trotz der genannten Herausforderungen einen robusten Grundschutz gegen bekannte Bedrohungen für die Privatheit dar und können einen wichtigen Baustein beim Schutz sensibler medizinischer Daten darstellen. Im Praxiseinsatz kann der Einsatz zusätzlicher Schutzmaßnahmen auf weiteren Ebenen angezeigt sein, wie in den folgenden Abschnitten erläutert wird.

Moderne Techniken der Datensynthetisierung haben das Potenzial, eine ausgewogenere Balance zwischen dem Schutz der Privatheit und der Nützlichkeit geteilter Daten zu erreichen. Dies gilt insbesondere, wenn Synthetisierungsprozesse mit umfassenden Schutzmodellen wie Differential Privacy kombiniert werden. Es bleibt jedoch auch mit diesen Methoden herausfordernd, sensible Informationen vor Offenlegung zu schützen und dabei gleichzeitig wissenschaftlich relevante, potenziell unbekannte Korrelationen in Daten zu erhalten. Obwohl synthetische Daten aufgrund ihrer künstlichen Natur nicht direkt re-identifizierbar sind und nicht durch Verknüpfung mit Identifizierenden oder identifizierten Daten angreifbar sind, besteht weiterhin die Herausforderung, Schutz vor Attribute und Membership Disclosure zu gewährleisten, ohne wissenschaftliche Erkenntnisse zu behindern. Zudem sind Fragen der Akzeptanz und Nutzbarkeit synthetischer Daten für wissenschaftliche Zwecke noch offen (siehe Dankar und El Emam[77] für eine Diskussion in Bezug auf Differential Privacy, das wie Synthetisierungsverfahren nicht-wahrheitserhaltende Daten generiert).

Zusammenfassend sind die Entwicklungen im Bereich der Anonymisierungsverfahren für medizinische Daten zwar vielversprechend, im praktischen Einsatz bestehen aber weiterhin Herausforderungen und rechtliche Unsicherheiten (siehe Abschnitte C.IV und C.V).

Chancen

- Die Hürden für den Datenzugang sind relativ niedrig.
- Forschende können die Daten auf gängige Weise nutzen und mit bekannten Werkzeugen und Methoden arbeiten.
- Die Anonymisierung der Daten als solche lässt sich nunmehr rechtssicher auf den Erlaubnistatbestand des § 6 Abs. 3 Satz 3 GDNG stützen. Eine Datenanonymisierung zu Forschungszwecken ist danach auch ohne Einwilligung der betroffenen Personen erlaubt, sofern die personenbezogenen Daten rechtmäßig gespeichert sind.

Risiken

- Die Anonymisierung von Daten ist komplex und es gibt keine klaren Richtlinien, die definieren, wann Daten als vollständig anonym gelten.
- Durch die Anonymisierung und Synthetisierung können wesentliche Informationen verloren gehen, was die Nützlichkeit der Daten für die Forschung stark beeinträchtigen könnte. Dies gilt insbesondere, aber nicht ausschließlich für hochdimensionale Daten

[76] El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One. 2011;6(12):e28071.
[77] Dankar FK, El Emam K. Practicing differential privacy in health care: A review. Trans. Data Priv. 2013; 6(1), 35-67.

wie Bilddaten und die klinisch und wissenschaftlich immer wichtiger werdenden OMICS-Daten (genetische Informationen können unter Erhalt ihres wissenschaftlichen Wertes kaum sinnvoll anonymisiert werden).

b) Szenario 2: Maßnahmen auf Datenebene in Kombination mit Bereitstellung in einer sicheren Verarbeitungsumgebung bei einer Datenzugangsstelle

Maßnahmen auf Prozessebene, wie sie in Szenario 2 und Szenario 3 beschrieben sind, können einen starken Schutz sensibler Daten vor Re-Identifizierung bieten. Jedoch schränken sie die Interaktionsmöglichkeiten mit den Daten ein, was die Nützlichkeit für wissenschaftliche Zwecke ebenfalls beeinträchtigen kann[78]. Methoden auf Prozessebene sind dennoch von großer Bedeutung, da sie in Kombination mit Maßnahmen auf Datenebene einen umfassenden Gesamtschutz bilden können.

Sichere Verarbeitungsumgebungen (auch Trusted Research Environments, TREs, genannt) basieren auf dem Prinzip, sensible Daten, auch aus unterschiedlichen Quellen, in einer geschützten Umgebung vorzuhalten und nur über speziell abgesicherte Zugangsmechanismen für die Auswertung zugänglich zu machen[79]. Die Verarbeitungsumgebungen können durch einen Datentreuhänder betrieben werden. Häufig werden dafür virtuelle Arbeitsumgebungen eingerichtet, die über das Internet erreichbar sind und in denen mit den Daten gearbeitet werden kann. Die Arbeit wird dabei protokolliert und mindestens stichprobenartig kontrolliert, um nicht nur die Risiken unbeabsichtigter Verletzungen der Privatsphäre zu minimieren, sondern auch vorsätzlichem Missbrauch aktiv vorzubeugen. Die Interaktionsmöglichkeiten mit den Daten können eingeschränkt sein. Beispiele umfassen den ”National Safe Haven“ des schottischen NHS[80] oder das ”Virtual Research Data Center” der Centers for Medicare & Medicaid Services der USA[81]. In Deutschland befindet sich das Forschungsdatenzentrum Gesundheit (FDZ) im Aufbau, das ebenfalls eine sichere Verarbeitungsumgebung bereitstellen soll.

Das Konzept von "**Datentreuhandstellen**" umfasst ein breites Spektrum an Ansätzen, die von unabhängigen Stellen, die Identitätsdaten pseudonymisieren und Einwilligungen verwalten, bis hin zu Betreibern von sicheren Verarbeitungsumgebungen oder gar Einrichtungen, die Datenanalysen im Auftrag durchführen, reichen[82]. Als unabhängige Instanzen im Rahmen einer Pseudonymisierung und Verknüpfung von Daten haben Treuhänder in der medizinischen Forschung eine lange Tradition und kommen umfangreich zum Einsatz. Als Einrichtung, die auch Primärdaten sicher verwahrt und durch abgesicherte Mechanismen zugänglich macht, haben sie erst in den vergangenen Jahren eine immer größere Bedeutung gewonnen und könnten in einem solchen Szenario als Zugangsstelle verwendet werden.

---

[78] Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021 Aug 12;21(1):242.
[79] Platt R, Lieu T. Data enclaves for sharing information derived from clinical and administrative data. JAMA. 2018;320:753–4. https://doi.org/10.1001/jama.2018.9342
[80] ISD Services. Use of the National Safe Haven [Internet]. Edinburgh: ISD Services. Verfügbar unter https://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/
[81] ResDAC. CMS Virtual Research Data Center (VRDC) [Internet]. Minneapolis: ResDAC. Verfügbar unter https://www.resdac.org/cms-virtual-research-data-center-vrdc
[82] Buchner B, Haber AC, Hahn HK, Prasser F, Kusch H, Sax U, Schmidt CO. Das Modell der Datentreuhand in der medizinischen Forschung. Datenschutz Datensicherheit-DuD. 2021;45:806-810.

Darauf hinzuweisen ist, dass auch in Szenario 2 Anonymisierungs- oder Synthetisierungsverfahren (vgl. Szenario 1) auf Datenebene eingesetzt werden können. Die dabei potenziell verbleibenden Restrisiken werden dann durch die sichere Verarbeitungsumgebung adressiert. Der Zugang zu den Daten wird ebenfalls über ein Antragsverfahren geregelt und setzt die Unterzeichnung einer Nutzungsvereinbarung voraus. Durch diese **Kombination von Maßnahmen auf Datenebene und Prozessebene** wird ein umfassender Schutz der Anonymität der Betroffenen erreicht, der aber möglicherweise rechtliche Anpassungen oder eine spezifische Rechtsgrundlage erfordert (vgl. Abschnitt C.V).

Chancen

- Durch die Kombination von Maßnahmen auf Daten- und Prozessebene kann ein hohes Schutzniveau bei gleichzeitigem Erhalt des wissenschaftlichen Wertes erreicht werden.
- Auch auf Datenebene nur schwer zu schützende Daten (wie Bilder und OMICS) können recht einfach aufgenommen werden.
- Die Daten verbleiben unter der Kontrolle einer vertrauenswürdigen und spezialisierten Einrichtung.

Risiken

- Es ist eine spezifische rechtliche Grundlage für die Verarbeitung pseudonymisierter Daten oder eine Erweiterung des Anonymitätsbegriffs auf die Nutzung in einem speziellen Kontext notwendig.[83]
- Da die Daten nur in sicheren Verarbeitungsumgebungen bereitgestellt werden, können Forschende nur dort verfügbare Werkzeuge und Methoden nutzen, mit denen sie unter Umständen nicht gut vertraut sind. Die Umsetzung anspruchsvoller und innovativer Vorhaben kann dann die Anpassung des zentral vorgehaltenen Werkzeugspektrums und der zugrundeliegenden Hardwareumgebung erfordern und damit neue Abhängigkeiten zwischen dezentraler Projektplanung und zentraler Infrastrukturvorhaltung induzieren.

c) Szenario 3: Dezentrale Bereitstellung durch föderiertes Lernen oder Analysieren

In diesem Szenario erfolgt die Datenverarbeitung über eine föderierte Plattform, die auch kryptographische Methoden einsetzen kann. Die Daten werden nicht direkt an Forschende weitergegeben, sondern bleiben innerhalb der teilnehmenden Institutionen. Durch den Einsatz verteilter statistischer Verfahren und verteilter Methoden des maschinellen Lernens werden die Daten so verarbeitet, dass die Anonymität gewahrt bleibt und keine Individualdaten die Institutionen verlassen. Forschende können Anfragen an die Plattform stellen und erhalten Ergebnisse ihrer Anfragen, ohne direkt auf die Rohdaten zugreifen zu können.

[83] Konkret zu letzterem Aspekt s.u. Punkt V am Ende.

Föderierte Datenanalysen oder föderiertes Machine Learning ermöglicht die Ausführung von Algorithmen auf dezentralisierten Daten, ohne dass die Daten ihre ursprünglichen Standorte verlassen. Dabei können mehrere Kommunikationsrunden stattfinden, bspw. um Modellgewichte iterativ zu optimieren[84]. Auch statistische Funktionen können auf verteilten Servern ausgeführt werden und statt Individualdaten aggregierte Ergebnisse zurückliefern[85]. Um den Austausch von Aggregatdaten oder Modellparametern in föderierten oder verteilten Szenarien abzusichern, können die in den Abschnitten C.III.1.a sowie C.IV genannten Methoden eingesetzt werden. Auch kryptographische Verfahren, bspw. aus dem Bereich der homomorphen Verschlüsselung, werden genutzt, da sie es ermöglichen, Berechnungen direkt auf verschlüsselten Daten durchzuführen[86].

Chancen

- Durch die Nutzung einer föderierten Plattform kann ein sehr hohes Schutzniveau erreicht werden.

- Dieses Modell bietet die Möglichkeit, innovative Ansätze der Datenbereitstellung zu entwickeln.

Risiken

- Die unterstützen Analysemöglichkeiten und -prozesse werden durch die Natur der föderierten Plattform und der verwendeten Methoden stark eingeschränkt.

- Die iterative Entwicklung, Verifikation und Validierung neuer Analyseverfahren sind je nach Grad der Interaktivität der föderierten Analyse-Architektur stark erschwert.

- Die Erkennung von Fehlern im Datenverarbeitungsprozess und Implausibilitäten auf Ergebnisebene ist ohne direkte Dateneinsicht deutlich herausfordernder, so dass die Wahrscheinlichkeit der Generierung unerkannt falscher Ergebnisse steigen kann.

- Gerade in Szenarien, wo Daten aus mehreren primär heterogenen Quellen stammen und nah an der Quelle föderiert verarbeitet werden, wird diese Herausforderung weiter exazerbiert, da z.B. standortspezifische Abweichungen in der Dokumentations- oder Harmonisierungspraxis leicht zu standortspezifischen inhaltlichen Fehlinterpretationen oder technischen Analysefehlern führen können, die für die für die Durchführung der föderierten Analyse Verantwortlichen nicht ohne weiteres erkennbar sind. Dezentrale Qualitätssicherungsmechanismen, die solche Phänomene zumindest partiell adressieren können, sind aufgrund des hohen Multiplikators für dezentrale Aufwände potenziell teuer - auch könnte die dauerhafte und breite Vorhaltung fachlich adäquat qualifizierter und erfahrener Mitarbeiter zur Besetzung der peripheren Verantwortlichkeiten für eine solche dezentrale Qualitätssicherung für das erwartbare methodologisch breite Nutzungsspektrum herausfordernd sein.

[84] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
[85] Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43(6):1929-44.
[86] Gentry C. A Fully Homomorphic Encryption Scheme. Stanford University; 2009

- Forschende müssen sich mit neuen Werkzeugen und Methoden vertraut machen.

d) Szenario 4: Umsetzung auf Basis einer breiten Forschungseinwilligung

In diesem mittel- bis langfristigen Szenario basiert die Datennutzung auf einer breiten Forschungseinwilligung, die von den Patientinnen und Patienten vor ihrem Intensivaufenthalt eingeholt wurde. Diese Einwilligung erlaubt es, die Daten der Patientinnen und Patienten in pseudonymisierter Form für Forschungszwecke zu nutzen und auch an Forschende herauszugeben. Forschende erhalten auf Antragsbasis Zugang zu diesen Daten.

Die unter C.I skizzierten Defizite des Datenschutzrechts in seiner gegenwärtigen Handhabe bringen es mit sich, dass Forschende regelmäßig auf die Einwilligung als Legitimationsgrundlage für eine Forschungsdatenverarbeitung zurückgreifen, insbesondere für die standortübergreifende, gemeinsame Nutzung von Daten von Patientinnen und Patienten bei der in bestimmten Fällen aus wissenschaftlich-technischen Gründen fast alternativlosen Datenzusammenführung, die für eine zweckändernde Nutzung von Daten von Patientinnen und Patienten für Forschungszwecke regelmäßig den Rückfall auf die breite Einwilligung (auch “Broad Consent") erfordert. Im Gegensatz zu studienspezifischen Einwilligungen, die z.B. im Falle der prospektiven Untersuchung risikobehafteter medizinischer Interventionen selbstverständlich weiterhin erforderlich bleiben, kann ein solcher “Broad Consent” früh und einmalig im Behandlungsablauf eingeholt werden und so die Untersuchung zahlreicher wissenschaftlicher Fragestellungen auf Grundlage der einmaligen Einholung einer Einwilligung ermöglichen – nicht nur ein ggf. erheblicher Effizienzgewinn, sondern bei geschickter Ausprägung potentiell auch eine Entlastung für die Patientinnen und Patienten, wenn die Breite der Einwilligung durch geeignete Kontroll- und Transparenzmaßnahmen begleitet wird[87]. Ein solches einwilligungsbasiertes Vorgehen (im Prinzip eine Form der “Opt In”-Datenspende) ist aber gerade für die für eine deutsche MIMIC relevante Population von Intensivpatientinnen und -patienten nur sehr eingeschränkt geeignet, da Intensivpatientinnen und -patienten aufgrund der Schwere der eine intensivmedizinische Behandlung erfordernden Erkrankungen oft schon zu Behandlungsbeginn nicht einwilligungsfähig sind (z.B. bei schwereren Formen des Schädel-Hirn-Traumas) und die eventuelle Wiedererlangung der Einwilligungsfähigkeit eng mit dem Krankheitsverlauf und der Erkrankungsschwere korreliert sind, so dass auf frühestens mit dem Behandlungsbeginn einholbare Broad Consent gründende Datenbestände eine für viele intensivmedizinisch relevante wissenschaftliche und qualitätssichernde Fragestellungen prohibitive systematische Selektionsverzerrung aufweisen dürften. Patientinnen und Patienten mit geringer Erkrankungsschwere und gutem Behandlungsergebnis würden so massiv überrepräsentiert, wohingegen Patientinnen und Patienten, die bei Behandlungsbeginn bewusstlos sind und das Bewusstsein nie wiedererlangen, gar nicht vertreten wären.

Aus datenschutzrechtlicher Warte gehen zudem mit der Einwilligung als Erlaubnistatbestand eine ganze Reihe von Wirksamkeitsvoraussetzungen einher, die seit jeher die Forschungs-

[87] Zenker S, Strech D, Ihrig K, Jahns R, Müller G, Schickhardt C, Schmidt G, Speer R, Winkler E, von Kielmansegg SG, Drepper J. Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new German national standard. J Biomed Inform. 2022 Jul;131:104096.

praxis vor besondere Herausforderungen stellen, wenn für eine Datenverarbeitung die Einwilligung als Legitimationsgrundlage herangezogen werden soll.[88] Dies beginnt schon mit dem Grundsatz der Informiertheit einer Einwilligung ("informed consent"). Angesichts der Komplexität der Datenverarbeitungsprozesse gerade im Forschungsbereich müssen die entsprechenden Informationen zwangsläufig detailliert und umfangreich ausfallen, um Rechtssicherheit für alle Beteiligten zu schaffen, und bergen damit stets das Risiko einer " Informationsüberflutung" und einer Überforderung von Patientinnen und Patienten. Auch im Fall des von der Medizininformatik-Initiative (MII) entwickelten Broad Consent muss der Einzelne immerhin sieben Seiten an Informationen und zusätzlich einen dreiseitigen Einwilligungstext lesen und verstehen - und das in einer (Behandlungs-)Situation, die vom einzelnen Betroffenen regelmäßig ohnehin bereits als belastend und überfordernd wahrgenommen wird.[89] Eine besondere Herausforderung ist des Weiteren auch die Wahrung der Freiwilligkeit der Einwilligung. Um dieser Herausforderung zu begegnen, war eine organisatorische Entkopplung der Einwilligung in den Broad Consent von für die Behandlung zwingenden Prozessschritten wie dem Abschluss des Behandlungsvertrages von Anfang an eine Forderung der begleitenden ethischen und datenschutzrechtlichen Beratung[90]; aktuell wird aktiv an der Etablierung eines national abgestimmten Vorgehens für diese Prozesse gearbeitet. So oder so ist aber die Organisation einer informierten, breiten Einwilligung in die Sekundärnutzung von Daten ein extrem ressourcenaufwändiger Vorgang, der nicht nur Institutionen, sondern auch Patientinnen und Patienten vor große Herausforderungen stellt. Nicht zuletzt stellt auch die freie und jederzeitige Widerrufbarkeit einer Einwilligung die Forschungspraxis vor erhebliche Herausforderungen, nicht zuletzt mit Blick auf die Notwendigkeit einer stabilen Datenbasis für die Forschung. Eine den rechtlichen Erfordernissen und Zusagen im Broad Consent genügende, ausreichend skalierbare Umsetzung von Widerrufsprozessen, die gleichzeitig auch mit einer praktikablen Forschungsdatennutzung vereinbar ist, induziert dementsprechend erhebliche organisatorische und technische Aufwände.

Chancen

- Durch die Einholung einer breiten Forschungseinwilligung können Daten in pseudonymisierter Form, also als personenbezogene Daten, Forschenden zugänglich gemacht werden.

- Forschende können die Daten in ihrer gewohnten Umgebung mit bekannten Werkzeugen und Methoden auswerten.

[88] Siehe zu alldem ausführlich auch schon Buchner, B. Forschungsdaten effektiver nutzen. Datenschutz Datensich 46, 555–560 (2022). https://doi.org/10.1007/s11623-022-1658-8.
[89] Siehe dazu auch Strech D, Graf von Kielmansegg S, Zenker S, Krawczak M, Semler S. Wissenschaftliches Gutachten „Datenspende" – Bedarf für die Forschung, ethische Bewertung, rechtliche, informationstechnologische und organisatorische Rahmenbedingungen. Erstellt für das Bundesministerium für Gesundheit, 30.03.2020; S. 100 f.: strukturelle Überforderung der Patienten, sowohl in situativer Hinsicht mit Blick auf die besondere Situation in der Klinik als auch inhaltlich mit Blick auf die Komplexität und die Ungewissheiten der Datenverarbeitungsprozesse.
[90] Zenker S, Strech D, Ihrig K, Jahns R, Müller G, Schickhardt C et al. Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new German national standard. J Biomed Inform. Juli 2022;131:104096.

- Aktuell liegt mit dem MII Broad Consent eine abgestimmte Lösung vor, deren breiterer Einsatz durch entsprechende Incentivierungsmechanismen schnell den rechtskonformen und relativ unkomplizierten Aufbau einer “deutschen MIMIC” für Teile der intensivmedizinischen Patientinnen- und Patientenpopulation ermöglichen könnten (z.B. elektives operatives Patientinnen- und Patientenkollektiv - Einwilligung hier z.B. im Rahmen der anästhesiologischen präoperativen Evaluation).

Risiken

∉ Die Implementierung dieses Szenarios erfordert eine umfangreiche Vorbereitungs- und Einwilligungsphase, die einige Zeit in Anspruch nimmt, bis eine ausreichend große Datenbasis zur Verfügung steht

∉ Die Umsetzung des Einwilligungsprozesses zu organisieren, zu dokumentieren, und alle abhängigen Prozesse wir z.B. den Widerruf zu organisieren und dauerhaft vorzuhalten, bedeutet nicht nur initial, sondern auch auf Dauer Zusatzaufwände, die nicht unmittelbar der Patientinnen- und Patientenversorgung oder dem Erkenntnisgewinn dienen. Dies ist nicht nur unter dem Aspekt der Finanzierung solcher Aufwände, sondern auch unter dem Aspekt der zunehmenden Knappheit von Personal potentiell problematisch.

∉ Insbesondere eine hochinteraktive Ausprägung von Consent- und Widerrufsprozessen, z.B. im Sinne einer sogenannten dynamischen Einwilligung, kann nicht nur die Selektionsverzerrung weiter verstärken, sondern auch selbst auf unterschiedliche Weise die Persönlichkeitsrechte der Patientinnen und Patienten gefährden.[91]

∉ Gerade für die intensivmedizinisch relevante Patientinnen- und Patientenpopulation erfordert eine adäquate Repräsentation der Gesamtpopulation letztlich die Einholung der Einwilligung von repräsentativen Bevölkerungsschichten, also eine Art Bürgerinnen- und Bürgerdatenspende[92], um die Selektionsverzerrung für bestimmte Untergruppen der Patientinnen- und Patientenpopulation in vertretbarem Rahmen zu halten (siehe Abschnitt C.II.2). Bestimmte Patientinnen- und Patientenkollektive könnten mit entsprechendem Ressourceneinsatz allerdings auch mit dem bisherigen MII Broad Consent-Modell sinnvoll erfasst werden, z.B. das Kollektiv der Patientinnen und Patienten, die sich einem elektiven (also in einem gewissen Rahmen im Voraus planbaren) operativen Eingriff unterziehen, wobei natürlich auch ein solches Vorgehen nicht frei von Selektionsverzerrungen wäre, da zum einen die Bereitschaft zur Einwilligung sicherlich in komplexer Art und Weise nicht zufällig z.B. mit dem aktuellen Gesundheitszustand und der Krankheitshistorie zusammenhängt, zum anderen aber auch trotz verfügbarer fremdsprachlicher Übersetzungen und Erläuterungen auch in einfacher Sprache durch etwaige Sprachbarrieren beeinflusst werden kann.

[91] Stellungnahme der AG Consent der Medizininformatik-Initiative zu patientenindividueller Datennutzungstransparenz und Dynamic Consent. Berlin: 2019; Verfügbar unter https://www.medizininformatik-initiative.de/sites/default/files/2019-09/MII_AG-Consent_Stellungnahme-Consent-Modelle_v05.pdf

[92] Strech D, Graf von Kielmansegg S, Zenker S, Krawczak M, Semler S. Wissenschaftliches Gutachten „Datenspende“ – Bedarf für die Forschung, ethische Bewertung, rechtliche, informationstechnologische und organisatorische Rahmenbedingungen. Erstellt für das Bundesministerium für Gesundheit, 30.03.2020

- ∉ In der datenschutzrechtlichen Diskussion ist die Wirksamkeit eines Broad Consent noch immer umstritten. Problematisiert wird vor allem die Informiertheit einer solchen Einwilligung.

2. Notwendige rechtliche Rahmenbedingungen

Welche rechtlichen Rahmenbedingungen für die Etablierung einer deutschen MIMIC gegeben sein müssen, hängt davon ab, welches der oben skizzierten Szenarien für die Umsetzung gewählt werden soll. Soll eine deutsche MIMIC im Wege der Datenanonymisierung ermöglicht werden (egal ob auf Daten- oder auf Prozessebene), bedarf es entscheidend einer rechtssicheren Definition der Anonymität von Daten. Unter welchen Voraussetzungen eine solche Definition überhaupt in Betracht kommt und wie diese im Einzelnen ausgestaltet sein könnte, ist Gegenstand der Ausführungen unter Punkt IV und V.

Wenn demgegenüber ein Umsetzungsszenario gewählt wird, bei dem Daten von Patientinnen und Patienten im Rahmen einer MIMIC nicht anonymisiert werden, sondern weiterhin einen Personenbezug aufweisen, bedarf es für die Datenverarbeitung einer entsprechenden rechtlichen Legitimationsgrundlage. Soweit diese im allgemeinen Datenschutzrecht oder im bereichsspezifischen Krankenhausrecht zu suchen ist, stellen sich die oben skizzierten Probleme eines restriktiven und zersplitterten datenschutzrechtlichen Regelwerks, in dessen Rahmen sich die Umsetzung einer MIMIC entweder gar nicht oder jedenfalls nicht rechtssicher umsetzen lässt. Mit dem Gesundheitsdatennutzungsgesetz wird zwar die Anonymisierung von Daten auf eine gesetzliche Grundlage gestellt, für die Weitergabe von personenbezogenen MIMIC-Daten an Dritte hilft aber auch dieses Gesetz nicht weiter.

a) Gesundheitsdatennutzungsgesetz (GDNG)

Am 14.12.2023 hat der Bundestag den Gesetzentwurf zum GDNG angenommen; am 2.2.2024 hat das Gesetz auch den Bundesrat passiert. § 6 GDNG regelt die Zulässigkeit einer Weiterverarbeitung von Versorgungsdaten zur Qualitätssicherung, zur Förderung der Patientinnen- und Patientensicherheit und zu Forschungszwecken. Die Vorschrift zielt auf ein "lernendes Gesundheitssystem" ab. Datenverarbeitende Gesundheitseinrichtungen sollen "die ihnen anvertrauten Informationen nicht bloß zum Zwecke der unmittelbaren Versorgung nutzen dürfen, sondern auch um aus ihnen zu lernen."[93] Verfolgt wird mit der Neuregelung also eben die Zielsetzung, die auch für eine MIMIC prägend ist.

Erleichtert wird durch § 6 GDNG der Aufbau einer MIMIC, soweit diese in ihrem Grundmodell darauf ausgelegt ist, Daten aus der klinischen Versorgung anonymisiert aufzubereiten, um sie dann auch Dritten zu Forschungszwecken zugänglich zu machen. Nach Auffassung des BfDI ist bereits eine entsprechende Anonymisierung der klinischen Daten eine erlaubnispflichtige zweckändernde Datenweiterverarbeitung, die nur auf Grundlage einer Einwilligung oder eines gesetzlichen Erlaubnistatbestands zulässig ist.[94] § 6 Abs. 3 Satz 3 GDNG liefert nunmehr eine

[93] Begründung zum Regierungsentwurf GDNG, BT-Drs. 20/9046, S. 54.
[94] BfDI Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche (Stand: 29.6.2020), S. 5.

solche rechtliche Grundlage. Die Vorschrift erlaubt eine Datenanonymisierung auch ohne Einwilligung der betroffenen Personen, sofern die personenbezogenen Daten rechtmäßig gespeichert sind. Damit soll es ausweislich der Entwurfsbegründung Forschungsverbünden ermöglicht werden, Daten von unterschiedlichen datenverarbeitenden Gesundheitseinrichtungen auszuwerten, „ohne dass für die betroffenen Personen ein weiteres Risiko entsteht."[95]

Für den weiteren Betrieb einer MIMIC, konkret für die Datenüberlassung *an Dritte* zu Forschungszwecken, ist die Neuregelung des § 6 GDNG hingegen ohne Relevanz. Können Daten im Rahmen einer MIMIC rechtssicher anonymisiert werden, unterliegt die weitere Datenverarbeitung ohnehin keinen (datenschutzrechtlichen) Grenzen mehr, es bedarf somit auch keiner rechtlichen Grundlage für eine Datenüberlassung an Dritte. Wird wiederum eine MIMIC-Variante gewählt, die keine Anonymisierung im rechtlichen Sinne gewährleistet, und handelt es sich damit bei der Preisgabe solcher MIMIC-Daten um eine erlaubnispflichtige Datenübermittlung, kann hierfür auch § 6 GDNG nicht als Erlaubnistatbestand herangezogen werden.

Zwar hat § 6 GNDG in seiner letzten Fassung nochmals eine Erweiterung erfahren und erlaubt nunmehr auch eine *Weitergabe* von Daten zu Forschungszwecken. Die Zulässigkeit einer solchen Weitergabe ist jedoch gemäß § 6 Abs. 3 Satz 4 GDNG auf einen Datenaustausch im Rahmen öffentlich geförderter Zusammenschlüsse von datenverarbeitenden Gesundheitseinrichtungen beschränkt. Beispielhaft spricht die Regelung Verbundforschungsvorhaben und Forschungspraxennetzwerke an. Zwar ist denkbar, dass im Falle einer multizentrisch verfassten MIMIC die beteiligten Kliniken ebenfalls unter § 6 Abs. 3 Satz 4 GDNG fallen und diese unter den dort genannten Voraussetzungen untereinander Daten austauschen dürfen. Insoweit handelt es sich jedoch lediglich um einen verbund*internen* Datenaustausch, während das eigentliche Spezifikum einer MIMIC, das Angebot einer auch *für Dritte* verfügbaren Datenbasis, von dieser Neuregelung auch in ihrer erweiterten Fassung nicht mehr gedeckt ist. Für eine Weitergabe von personenbezogenen Daten an Dritte außerhalb eines Forschungsverbunds bleibt es vielmehr bei der Grundregel des § 6 Abs. 3 S. 2 GDNG, dass entweder die betroffene Person eingewilligt haben oder aber eine anderweitiger gesetzlicher Erlaubnistatbestand einschlägig sein muss.[96]

b) § 363 SGB V und EHDS

Im Zuge der Verabschiedung des Gesundheitsdatennutzungsgesetzes (GDNG) ist auch die Vorschrift des § 363 SGB V dahingehend geändert werden, dass künftig Daten aus der elektronischen Patientinnen- und Patientenakte auf Grundlage eines Opt-Out-Verfahrens für Forschungszwecke verarbeitet werden dürfen. Durch die automatisierte Bereitstellung einzelner strukturierter Datenkategorien an das FDZ soll die Datennutzung für Forschungszwecke erheblich forciert und vereinfacht werden. Zum Schutz der Daten wird eine zweifache Pseudonymisierung (in der elektronischen Patientinnen- und Patientenakte, ePA, selbst und in der Vertrauensstelle) angewendet.[97]

---

[95] Begründung zum Regierungsentwurf GDNG, BT-Drs. 20/9046, S. 54.

[96] Zum anderen setzt § 6 GDNG auch der Möglichkeit einer Datenverknüpfung sehr enge Grenzen. Die Entwurfsbegründung stellt insoweit klar, dass § 6 keine Legitimationsgrundlage dafür liefert, im Sinne einer breiteren und aussagekräftigeren Datenbasis die Daten aus dem Versorgungskontext um andere personenbezogene Daten zu ergänzen, die nicht unmittelbar für den Versorgungsanlass benötigt werden; Begründung zum Regierungsentwurf GDNG, BT-Drs. 20/9046, S. 55.

[97] Begründung zum Regierungsentwurf, BT-Drs. 20/9046, S. 72.

Perspektivisch ist es durchaus denkbar, mit einer entsprechend ertüchtigten ePA eine Art von "Versorgungs-MIMIC" im Rahmen der ePA aufzubauen. Eine entsprechende Forschungsnutzung von ePA-Daten kann aber selbstverständlich nur solche Daten umfassen, die in geeigneter Art und Weise in die ePA übertragen werden. Insofern teilt dieser Ansatz sämtliche proximalen strukturellen Voraussetzungen alternativer Ansätze zur feingranularen Datenverfügbarmachung beginnend bei der Erschließung, Harmonisierung und Integration der Quellen, die dann systematisch in großen Maßstab in die hierfür infrastrukturell zu ertüchtigende ePA ausgeleitet werden müssten und dann über das ebenfalls entsprechend einzurichtende FDZ nutzbar gemacht würden.

Gleiches gilt dann auch für mögliche Datennutzungsszenarien, wenn im Zuge der geplanten Errichtung eines gemeinsamen europäischen Gesundheitsdatenraums (European Health Data Space, EHDS) elektronische Gesundheitsdaten in weitem Umfang auch für eine Sekundärnutzung zur Verfügung gestellt werden, insbesondere auch für Zwecke der wissenschaftlichen Forschung im Gesundheitssektor (Art. 34 Abs. 1 lit. e EHDS-VOE).[98] Dazu sollen bei den Mitgliedstaaten sog. Zugangsstellen für Gesundheitsdaten eingerichtet, die den Datentransfer vom jeweiligen Dateninhaber zu Nutzungsberechtigten organisieren sollen.

c) Registergesetz

Ausgehend von der Definition eines Registers als „ein organisiertes System, in welchem basierend auf einer zuvor festgelegten Fragestellung standardisiert Daten von Beobachtungseinheiten erhoben werden",[99] zählt auch eine MIMIC zu den medizinischen Registern, die durch ein neues Registergesetz einen einheitlichen Regelungsrahmen erhalten sollen. Zielsetzung eines solchen Registergesetzes ist unter anderem auch die Schaffung bundeseinheitlicher Regelungen für die Erhebung und Verarbeitung von Registerdaten, um der Rechtszersplitterung abzuhelfen, die im Registerbereich nochmals besonders ausgeprägt ist.

Perspektivisch wird für bestimmte Register die Möglichkeit einer Datenerhebung ohne Einwilligung der betroffenen Patientinnen und Patienten anvisiert, stattdessen soll insoweit ein Recht zum Widerspruch eingeräumt werden.[100] Gleiches wird dann auch für die Zulässigkeit einer Bereitstellung der Registerdaten für Dritte diskutiert.[101] Ob und inwieweit auch eine MIMIC unter dieses Registerprivileg fallen könnte, ist nach derzeitigem Stand nicht abzusehen. Von Relevanz wäre ein solches Registerprivileg für eine MIMIC jedenfalls dann, wenn eine solche MIMIC keine (rechtssicher) anonymisierten Daten, sondern personenbezogene Daten an Dritte übermittelt.

---

[98] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über den Europäischen Raum für Gesundheitsdaten in der vom Parlament abgeänderten Fassung vom 13. Dezember 2023.

[99] Gutachten zur Weiterentwicklung medizinischer Register zur Verbesserung der Dateneinspeisung und -anschlussfähigkeit. 2021 Verfügbar unter: https://www.bundesgesundheitsministerium.de/ fileadmin/Dateien/5_Publikationen/Gesundheit/Berichte/REG-GUT-2021_Registergutachten_BQS-TMF-Gutachtenteam_2021-10-29.pdf.

[100] Algermissen M. Auf dem Weg zu einem Registergesetz 8. Verfügbar unter: https://www.tmf-ev.de/sites/default/files/2023-07/news-algermissen-bmg-auf-dem-weg-zu-einem-registergesetz-registertage-2023.pdf.

[101] Siehe insoweit etwa die Forderung nach einem allgemeinen, voraussetzungslosen Widerspruchsrecht für den Fall, dass eine Übermittlung von Registerdaten an Dritte unter Verzicht auf eine vorherige Einbindung der betroffenen Person erlaubt werden sollte; Datenschutzkonferenz (DSK). Rahmenbedingungen und Empfehlungen für die gesetzliche Regulierung medizinischer Register. 22./23.11.2023.

d) Rechtsvereinheitlichung

Mangels spezifischer Erlaubnistatbestände für eine „deutsche“ MIMIC ist die Frage, ob und unter welchen Voraussetzungen MIMIC-Daten an Dritte zu Forschungszwecken herausgegeben werden dürfen, auf Grundlage des allgemeinen Datenschutzrechts bzw. des vorrangigen Landeskrankenhausrechts zu beantworten. Insoweit bedarf es dann aber zunächst einmal – ebenso wie auch für andere Arten der Forschungsdatenverarbeitung – dringend einer Rechtsvereinheitlichung auf Ebene der landesrechtlichen Vorschriften (s. zu dieser Problematik schon ausführlich oben C.I.2). Bislang fehlt es an länderübergreifend einheitlichen Vorgaben für die Forschungsdatenverarbeitung, die rechtssicher zu handhaben wären und dabei den großzügigen Privilegierungsrahmen, wie ihn die DS-GVO für die Forschungsdatenverarbeitung erlaubt, so weit wie möglich ausschöpfen.

Erst vor kurzem hat auch die Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (Datenschutzkonferenz – DSK) in ihrer Entschließung vom 23.11.2023 den Bundes- und die Landesgesetzgeber aufgefordert, im Sinne der Forschung für eine Rechtsvereinheitlichung im Datenschutzrecht zu sorgen. Vorschläge, wie eine solche Rechtsvereinheitlichung erreicht werden könnte, sind schon zuvor gemacht worden, etwa in Form einer Regelung auf Bundesebene in inhaltlicher Anlehnung an § 27 BDSG,[102] eines Bund-Länder-Staatsvertrags[103] oder eines Forschungsgesetzes auf Bundesebene für diejenigen Rechtsbereiche, die der Gesetzgebungszuständigkeit des Bundes unterfallen, flankiert durch eine „Einladung“ an die Bundesländer, ihr Landesrecht ebenfalls entsprechend mittels dynamischer Verweisungsnormen anzupassen.[104]

Jedenfalls auf Landesebene scheinen bislang aber der gesetzgeberische Wille und/oder die Fähigkeit zu einer Rechtsvereinheitlichung kaum oder gar nicht vorhanden zu sein. Was wiederum Initiativen zur Rechtsvereinheitlichung auf Bundesebene angeht, sind diese zwar mittlerweile zu verzeichnen, scheitern aber mitunter bereits an strittigen Kompetenzfragen. Exemplarisch dafür sind die Versuche des Bundesgesetzgebers, für länderübergreifende Forschungsvorhaben die Forschungsklausel des § 27 BDSG als einheitlichen Rechtsstandard zu etablieren. Der erste Versuch, diesen Regelungsansatz auf die Bundeskompetenz im sozialrechtlichen Bereich zu stützen, ist inzwischen vom Bundesgesetzgeber selbst als erfolglos verbucht worden: Die entsprechende Norm (§ 278a SGB V) sei „bislang oft zu eng ausgelegt und entsprechend selten angewandt“ worden.[105] In einem zweiten Versuch sollte daher die Regelung des § 287a SGB V inhaltlich in das Gesundheitsdatennutzungsgesetz überführt werden. § 3 Abs. 2 des Referentenentwurfs sah entsprechend vor, dass für länderübergreifende Forschungsvorhaben § 27 BDSG anzuwenden sei.[106] Auch an diesem Regelungsvor-

[102] Dierks C, Kircher P, Engelke K, Haase M. Lösungsvorschläge für ein neues Gesundheitsforschungsdatenschutzrecht in Bund und Ländern. 15.09.2019 (S. 100 ff.).
[103] Weichert T, Krawczak M., Vorschlag einer modernen Dateninfrastruktur für die medizinische Forschung in Deutschland. MIBE 2019, Vol. 15(1), 6; ebenso Weichert T. Die Forschungsprivilegierung in der DS-GVO. Zeitschrift für Datenschutz 2020 (1), 18.
[104] Bernhardt U, Ruhmann I, Weichert T. Plädoyer für ein medizinisches Forschungsgesetz. 22.02.2021. Verfügbar unter: https://www.netzwerk-datenschutzexpertise.de/sites/default/files/gut_2021_02_medforschungdatens_final.pdf.
[105] Regierungsentwurf zum GDNG vom 1.11.2023, BT-Drs. 20/9046, S. 53.
[106] Referentenentwurf zum GDNG vom 4.8.2023, S. 9.

schlag entzündete sich dann allerdings wieder Kritik wegen zweifelhafter Gesetzgebungskompetenz des Bundes im Krankenhausbereich, nicht zuletzt von Seiten der DSK in ihrer Stellungnahme zum GDNG-Entwurf vom 14.8.2023. Letztlich hat die Regelung des § 287a SGB V mit ihrem Verweis auf § 27 BDSG in die finale Fassung des GDNG auch keinen Eingang mehr gefunden.

Dabei lässt sich die Gesetzgebungskompetenz des Bundes für eine einheitliche Regulierung des Forschungsdatenschutzrechts durchaus begründen – und zwar so wie es in der Begründung zum Referentenentwurf des GDNG bereits angesprochen und an anderer Stelle auch schon ausführlicher begründet worden ist, nämlich mit Art. 74 Abs. 1 Nr. 13 Grundgesetz (GG, Förderung der wissenschaftlichen Forschung).[107] Mit Blick auf die disparaten Lösungsansätze zum Forschungsdatenschutz, nicht nur im Landeskrankenhausrecht, sondern auch im allgemeinen Landesdatenschutzrecht, ist es evident, dass sich länderübergreifende medizinische Forschung in erheblichem Maße mit binnenstaatlichen (Rechts-)Hindernissen konfrontiert sehen und deshalb im Sinne des Art. 72 Abs. 2 GG „die Wahrung der Rechts- oder Wirtschaftseinheit im gesamtstaatlichen Interesse eine bundesgesetzliche Regelung erforderlich macht".[108]

Ein erster Schritt hin zu einheitlichen rechtlichen Rahmenbedingungen für länderübergreifende Forschungsprojekte ist zumindest aber mit dem neuen § 6 Abs. 3 S. 4 GDNG gemacht worden, auf dessen Grundlage eine gemeinsame Datennutzung und -verarbeitung im Rahmen der Verbundforschung und vergleichbarer Zusammenschlüsse ermöglicht wird. Für ein Teilen von Forschungsdaten nach dem Konzept einer MIMIC kann diese Vorschrift allerdings nicht herangezogen werden (ausführlicher dazu schon oben III.2.a).

3. Weitere Einflussfaktoren

a) Einflussmöglichkeiten der Bundesbehörden

Ein konzertiertes Zusammenwirken wesentlicher Bundesbehörden und Institutionen des deutschen Gesundheitswesens könnte erheblich zu einer effektiven und effizienten Schaffung zentraler Voraussetzungen für die Umsetzung einer wirksamen und international kompetitiven Ausprägung einer deutschen MIMIC beitragen. Wichtige Akteure umfassen das Institut für das Entgeltsystem im Krankenhaus (InEK), das Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (IQTiG), das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), das Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM), der BfDI sowie das Robert Koch-Institut (RKI).

[107] Referentenentwurf zum GDNG vom 4.8.2023, S. 21 sowie ausführlicher zur Gesetzgebungskompetenz des Bundes im Bereich der Forschungsdatenverarbeitung Dierks C, Kircher P, Engelke K, Haase M. Lösungsvorschläge für ein neues Gesundheitsforschungsdatenschutzrecht in Bund und Ländern. 15.09.2019 (S. 35 f., 94).

[108] Dierks C, Kircher P, Engelke K, Haase M. Lösungsvorschläge für ein neues Gesundheitsforschungsdatenschutzrecht in Bund und Ländern. 15.09.2019 (S. 35 f., 96 ff.); tendenziell großzügig auch Strech D, Graf von Kielmansegg S, Zenker S, Krawczak M, Semler S. Wissenschaftliches Gutachten „Datenspende" – Bedarf für die Forschung, ethische Bewertung, rechtliche, informationstechnologische und organisatorische Rahmenbedingungen. Erstellt für das Bundesministerium für Gesundheit, 30.03.2020 (S. 122).

Zu den offensichtlichen und aktuell in Planung oder z.T. schon in Umsetzung befindlichen Handlungsfelder zählt hier insbesondere das Angebot (ggf. in Zusammenarbeit mit externen Dienstleistern) von Treuhandstellendienstleistungen zur Gewährleistung des Schutzes direkt identifizierender Daten von Patientinnen und Patienten. Durch die in modernen Datenschutzkonzepten strukturell verankerte und technisch durchzusetzende informationelle Gewaltenteilung könnten Bundesinstitute wie das RKI im Rahmen ihrer hoheitlichen Aufgaben die Prozessierung der direkt identifizierenden Daten verantworten, ohne hierdurch gleich Zugang zu z.B. den genetischen Daten aller Patientinnen und Patienten zu erhalten. Weiterhin wichtig wäre die systemweite Orchestrierung von Record Linkage-Verfahren, die es ermöglichen, der gleichen Person zugehörige Daten bei der Zusammenführung aus unterschiedlichen Quellen systematisch zuzuordnen, um z.B. Mehrfachzählungen seltener Ereignisse oder Erkrankungen zu verhindern. Gleiches gilt für die systematische Organisation der Anbindung bestehender und in Entwicklung befindlicher Datennutzungsinfrastrukturen an internationale Vorhaben wie den EHDS. Hierbei handelt es sich jeweils um Aufgaben, zu denen u.a. das BfArM mit dem FDZ beitragen kann.

Auch in IQTiG und IQWiG sind bereits wichtige Strukturen angelegt, die mittels der vorgeschlagenen Strukturentwicklung bzgl. dynamischer, risikoadjustierter Qualitätsmessung und systematischer Evaluation neuer diagnostischer und therapeutischer Verfahren in ihrer Rolle als “Sensor” im lernenden Gesundheitssystem weiter gestärkt werden könnten. Das InEK wiederum kann über Erlösanreize einen wichtigen Aktuator zur Implementierung der geschlossenen Rückkopplungsschleife bilden. Die geplante Digitalagentur als Nachfolgeorganisation der gematik unter 100%ger Kontrolle des Bundesministeriums für Gesundheit (BMG) schließlich kann durch Vorantreiben eines fachlich, nicht kommerziell, koordinierten und getriebenen Standardisierungsprozesses für medizinische Dokumentationsinhalte und deren eindeutiger Repräsentation das entscheidende, agil weiterzuentwickelnde Fundament für diese Strukturentwicklung schaffen.

Zentral und im Rahmen dieses Gutachtens nicht eindeutig zu beantworten ist die Frage der übergeordneten Orchestrierung all dieser komplexen und voneinander abhängigen Aktivitäten. Abzuraten ist hier von der Verortung innerhalb einer Bundesoberbehörde, da diese letztlich zumindest informell auch untereinander in Verantwortungs- und Ressourcenkonkurrenz stehen können. Strukturell sinnhaft erscheint hier eher die Verortung der Gesamtkoordinationsverantwortung für diese Aktivitäten in einer Einheit oder einem Gremium, dass von solchen Wettbewerbskonstellationen weitestmöglich entkoppelt ist und idealerweise auch ein konzertiertes Vorgehen über die Ressortgrenzen der Bundesministerien hinweg sicherstellt. Jedenfalls könnte eine strukturelle Verankerung einer strategisch orientierten Abstimmung und Orchestrierung der oft durch das Bundesministerium für Bildung und Forschung geförderten Forschungs-, Entwicklungs- und Innovationsvorhaben mit den eher versorgungsorientierten Aktivitäten des BMG (sowie in Spezialfällen auch des Bundesministeriums der Verteidigung und des Bundesministeriums für Ernährung und Landwirtschaft sowie des Bundesministeriums für Wirtschaft und Klimaschutz) einen sehr großen Mehrwert schaffen. Dies gilt einerseits mit Blick auf die Vermeidung von Redundanzen und die Maximierung von Synergien, anderseits aber auch mit Blick auf die noch bessere strategische Ansteuerung des Forschungs- und Innovationsprozesses entlang der absehbaren Bedarfe der Gesundheitsversorgung.

b) Datenaufbereitung

Neben der Schaffung der im vorherigen Abschnitt skizzierten notwendigen rechtlichen Rahmenbedingungen für ein ausgewähltes Umsetzungsszenario, ist ein zentraler Aspekt die Förderung der Generierung und Vergleichbarkeit nutzbarer Daten. Dies beinhaltet die Implementierung einer standardisierten Dokumentation in der klinischen Versorgung, welche nicht separat durch Dokumentare erfolgen sollte, sondern integraler Bestandteil des Versorgungsprozesses sein muss. Eine wichtige Rolle spielt hierbei die Entwicklung einer interoperablen Datenrepräsentation (koordiniert durch Interop Council[109] bzw. der Standard für Informationstechnische Systeme in Krankenhäusern, ISiK), basierend auf bestehenden Datenmodellen, wie dem Erweiterungsmodul Intensivmedizin des Kerndatensatzes[110] der MII. Dies sollte in enger Abstimmung mit relevanten Organisationen wie der Deutschen Gesellschaft für Anästhesiologie und Intensivmedizin (DGAI) und der Deutschen Interdisziplinären Vereinigung für Intensiv- und Notfallmedizin (DIVI) schrittweise über alle Versorgungsstufen hinweg umgesetzt werden. Parallel dazu sollte die Förderung einer robusten Infrastruktur zur Erfassung von Biosignaldaten vorangetrieben werden, Für eine detailliertere Analyse verweisen wir auf Abschnitt C.III.4.b.

Im nächsten Schritt müssen die vergleichbaren und interoperablen Daten für die Sekundärnutzung für qualitätsgesicherte und genehmigte Forschungszwecke zugänglich gemacht werden. Entsprechende Umsetzungsoptionen wurden im vorherigen Abschnitt skizziert. Dabei ist es essentiell, den Sekundärnutzungsprozess transparent zu gestalten, seine faire Ausprägung zu gewährleisten und den kompletten Nutzungszyklus sensibler Daten engmaschig zu überwachen. Dazu gehört auch die Einführung von Strafmaßnahmen bei Missbrauch solcher Daten oder aus diesen Daten abgeleiteter Informationen. Hierbei sollte zwischen Angriffen durch nicht am Datennutzungsprozess beteiligte Personen, für die vermutlich das bestehende Recht ausreicht, und vermeidbarem Disclosure sowie missbräuchlicher Nutzung durch am Datennutzungsprozess beteiligte Personen unterschieden werden. Für letztere könnte eine Strafbewehrung in Betracht gezogen werden, wie sie nun auch Eingang in das GDNG gefunden hat. § 7 GDNG normiert eine Reihe von Geheimhaltungspflichten, um einen Missbrauch von zu wissenschaftlichen Forschungszwecken überlassenen Daten zu anderen Zwecken auszuschließen. Wird gegen diese Geheimhaltungspflichten verstoßen, sieht § 9 GDNG entsprechende Strafvorschriften vor. Für eine detailliertere Analyse der Umsetzungsvarianten verweisen wir auch hier auf Abschnitt C.III.4.b.

4. Ressourcenbedarfe für eine Umsetzung

Der erforderliche Ressourceneinsatz hängt stark mit dem geplanten Vorgehen zusammen. Um unterschiedliche Umsetzungsvarianten vergleichen zu können, versuchen wir im Folgenden zunächst eine Analyse der strukturellen Voraussetzungen der erfolgreichen MIMIC-Umsetzung und deren unterschiedlich aufwändigen Ausprägungen bzgl. der unterschiedlichen, mindestens partiell orthogonalen Handlungsfelder, die Aufwand und Erreichbares determinieren. Höherer Aufwand bedeutet in der Regel auch einen höheren Zielerreichungsgrad. Sämtlich im Folgenden beschriebenen Strukturvoraussetzungen gelten übrigens auch weitestge-

[109] https://www.ina.gematik.de/mitwirken/expertengremium
[110] https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative

hend generisch für eine Ausweitung der Ziele des Vorhabens über die Grenzen der Intensivmedizin hinaus und könnten in den aufwändigeren Ausprägungen als Modell für einen weiteren systematischen Ausbau des deutschen Gesundheitssystems zu einem datengetrieben lernenden System dienen.

Nach Analyse der die Aufwände maßgeblich beeinflussenden strukturellen Voraussetzungen versuchen wir im abschließenden Abschnitt eine Synthese der verschiedenen Handlungsoptionen, um hieraus Umsetzungsempfehlungen abzuleiten.

Die angesetzten Ressourcenbedarfe sind als grobe erste Abschätzung zu verstehen. Eine Umsetzungsplanung, die eine belastbarere Ressourcenplanung erlaubt, wäre nach politischer Festlegung der gewünschten Umsetzungsvariante(n) bzw. der grundsätzlichen Stoßrichtung allerdings relativ zeitnah durch eine geeignete, interdisziplinär besetzte Planungsgruppe möglich. Wir schätzen die Planungsaufwände für eine ausschreibungsreif belastbare Aufwandsschätzung und Umsetzungsplanung auf zwischen. 0,5 und 2 Personenjahren, abhängig vom Anspruch und der Gesamtkomplexität der angestrebten Umsetzung.

a) Strukturelle Voraussetzungen einer erfolgreichen Umsetzung

aa) Standardisierung der Dokumentationsinhalte, interoperable Beschreibung der Datenelemente und ihres Kontextes und Umsetzung dieser Standards in der klinischen Routine

Ebenso wie aktuell auf europäischer Ebene im Zuge des EHDS mittels Standardisierung auf eine *grenzüberschreitende* Interoperabilität und Übertragbarkeit von Daten hingewirkt werden soll, um die Kontinuität der Gesundheitsversorgung und effiziente Gesundheitssysteme zu gewährleisten,[111] erfordert auch im nationalen Rahmen eine optimale gemeinsame Datennutzung aus unterschiedlichen Quellen bei unterschiedlichen Leistungserbringern die Harmonisierung der Datenerfassung. Für mittels Medizingeräten durchgeführte Messungen gibt es hier über vorhandene oder in Entwicklung befindliche Standards[112] bereits technisch und organisatorisch konkret definierbare Vorgehensmodelle. Allerdings steht die tatsächliche Umsetzung schon in Ermangelung der basalen Infrastruktur an vielen Standorten noch weitestgehend aus. Ein breiter Einsatz des Erweiterungsmodules Intensivmedizin des Kerndatensatzes der MII könnte hier einen ersten Durchbruch bedeuten, sofern dessen tatsächliche Umsetzung an den Standorten mit ausreichend Ressourcen unterfüttert und durchgesetzt werden könnte.

Der Nutzen der Gesamtstruktur hängt unmittelbar mit der inhaltlichen und technischen Abdeckungsbreite der standardisierten Dokumentationselemente ab. Hier sind diverse Umsetzungsstufen vorzufinden: Dies beginnt mit bereits weitestgehend standardisierten oder leicht standardisierbaren Datenelementen wie physiologische Messwerte (Herzfrequenz, Blutdruck) und Laborwerte. Andere Daten wie Medikationsdaten sind zumindest gut strukturiert, aber

[111] S. insb. Art. 6 (Europäisches Austauschformat für elektronische Patientenakten) und Art. 23 (Gemeinsame Spezifikationen) des Vorschlags für eine Verordnung des Europäischen Parlaments und des Rates über den Europäischen Raum für Gesundheitsdaten in der vom Parlament abgeänderten Fassung vom 13. Dezember 2023.
[112] DIN EN ISO 11073-10101:2021-01 Medizinische Informatik - Geräteinteroperabilität - Teil 10101: Kommunikation patientennaher medizinischer Geräte; IEEE 11073 Service-oriented Device Connectivity (SDC)

noch nicht vollständig interoperabel repräsentiert. Schließlich gibt es Daten, bzgl. deren Erhebung und typischer Ausgestaltung oder gar Strukturierung noch gar kein übergreifender Konsens besteht. Gerade in letzteren Datenarten werden sich aber regelmäßig viele für die Systemsteuerung und wissenschaftliche Analyse extrem wichtige Informationen zum Behandlungsverlauf finden, so z.B. der Zeitpunkt der erstmaligen Erwägung oder Stellung einer Verdachtsdiagnose oder anderen klinischen Hypothese oder der Evaluation und Anpassung der Therapieziele. Die Entwicklung und breite Durchsetzung einer praxisgerechten strukturierten, standardisierten und interoperablen Abbildung klinischer Versorgungsprozesse erfordert allerdings einen aufwändigen Konsens- und Entwicklungsprozess. Die Realisierung eines solchen Prozesses würde aber potentiell nicht nur einen massiven Mehrwert für die hier primär betrachteten Sekundärnutzungsszenarien entfalten, sondern würde auch eine interoperable Basis für klinische Entscheidungsunterstützungsverfahren schaffen. Zudem könnte dieser Prozess auch erheblich zur mitarbeiter- und patientinnen- und patientenzentrierten Weiterentwicklung der Abrechnungssystematik und der damit zusammenhängenden anreizbasierten Steuerung des Gesundheitssystems beitragen. U.a. würde so nicht nur eine solide Grundlage für eine systematische Incentivierung der Prozessqualität, sondern potentiell auch für einen datengetriebenen, auf risikoadjustierte Ergebnisqualität abzielenden ergebnisorientierten Vergütungsansatz geschaffen. Hierdurch wären die Umsetzungsaufwände zumindest partiell gegenfinanzierbar und besser begründbar. Drei wesentliche Umsetzungsvarianten für die standardisierte und interoperable Datenbeschreibung und -erfassung sind möglich:

- **Variante (a) Minimaldatensatz:** Nur bereits existente standardisierte Datenelemente (oft primär erlössicherungsbezogen), Labordaten sowie medizintechnische Daten (mit oder ohne hochauflösende Biosignaldaten) werden primär erschlossen. Die breite Implementierung wird durch Schaffung adäquater Anreizsysteme unterstützt, die praxisgerechte Umsetzung und Weiterentwicklung der interoperablen Spezifikationen wird durch Personalressourcenallokation für die systematische Weiterentwicklung in Höhe weniger Vollzeitäquivalente für Deutschland insgesamt auf eine nachhaltige Basis gestellt. Dieses “Datenkümmererteam” müsste aber durch geeignet hoch- und mehrfach qualifizierte Personal besetzt werden, aktuell eine zentrale Herausforderung, deren Lösung auch durch eine Erhöhung der Sichtbarkeit und offensichtlichen Relevanz solcher Aktivitäten unterstützt werden könnte. Es braucht engagierte, klinisch und technisch maximal qualifizierte ärztliche “Datenkümmerer”, für die diese Spezialisierung einen plausiblen Karrierepfad eröffnet, da hier eine substantielle Überlappung mit dem Talentpool besteht, aus dem sich auch die medizinische Wissenschaft und die klinische Führungselite speist. Nicht berücksichtigt sind hier die basalen Infrastrukturaufwände z.B. für die Beschaffung der notwendigen Integrationskomponenten, um Biosignale aus dem Patientinnen- und Patientenmonitoring dauerhaft speichern und vorhalten zu können, da diese (anders als Bilddaten) heute regelmäßig nur temporär gespeichert und dann verworfen werden. Damit geht aber - jeden Tag - ein großer Datenschatz verloren geht. Um diese Daten breiter verfügbar zu machen, wäre eine Finanzierung oder Incentivierung der Schaffung von Biosignal-Integrations- und Archivierungsinfrastrukturen notwendig.

- **Variante (b) Datensatz mit mittlerer Breite:** Zusätzlich werden Medikationsdaten und ggf. detaillierte Informationen zur Bilanz von Einfuhr und Ausfuhr von behandlungsrelevanten Substanzklassen wie z.B. kristallinen Flüssigkeiten, Elektrolyten, Ei-

weißen und anderen Blutbestandteilen (sog. Bilanzdaten) sowie weitere niederschwellig standardisierbare Dokumentationsinhalte, die keinen aufwändigen Standardisierungsprozess erfordern, erschlossen. Im Falle der Realisierung geeigneter Anreizsysteme für die Leistungserbringer (und damit sekundär die Softwarehersteller) kann dies strukturell analog zur Variante a) angegangen werden. Die Ressourcenausstattung für die Spezifikationsaktivitäten sowie die Eignung und Struktur der Anreizsysteme beeinflussen die Umsetzungsgeschwindigkeit maßgeblich, das Aufgabenspektrum der "Datenkümmerer" wird breiter und anspruchsvoller. Für die für eine zukunftssichere Aufstellung der Strukturen essentielle Integration von Bilddaten und OMICS-Daten sind zusätzliche Investitionen in Integrations-, Datenhaltungs- und Analyseinfrastrukturen sowie zusätzlich Spezifikationsaktivitäten auf Metadatenebene erforderlich.

- **Variante (c) Umfangreicher Datensatz:** Es werden in zunehmendem Maße strukturierte und standardisierte klinische Dokumentationsverfahren konsentiert und breit ausgerollt. Anders als bei den Varianten a und b ist hier die Organisation eines breit klinisch mitgetragenen, aber verbindlichen und dauerhaften Konsentierungsprozesses erforderlich, da die klinische Dokumentation sich selbstverständlich kontinuierlich dem Wandel der medizinischen Praxis anpassen muss. Auch hier wären Anreizsysteme für die Leistungserbringer sinnvoll, die für eine breite und stets aktuelle Umsetzung der nationalen Standards sorgen. Additiv zu den Aufwänden für Varianten a) und b) müssten aber auch dauerhafte nationale Konsensprozesse organisiert werden, die dann ein essentieller und zentraler Bestandteil des lernenden Gesundheitssystems der Zukunft würden. Umsetzbar ist dies am ehesten auf Basis einer Weiterentwicklung und zweckbezogenen Integration existierender Strukturen aus Interop Council, der MII, dem Netzwerk Universitätsmedizin (NUM) und Fachgesellschaften in eine gemeinsame Governance-Struktur, die dann *einen* Versorgungs- und Sekundärnutzungsstandard definiert und konsentiert. Das existierende und in kontinuierlicher Weiterentwicklung befindliche Erweiterungsmodul "Intensivmedizin" eignet sich hierfür als solide Grundlage, die viele klinisch relevante Aspekte bereits heute abbildet. Die Schaffung eines Pools engagierter und entsprechend qualifizierter "Kümmerer" durch eine attraktive Ausgestaltung dieser Prozesse und der assoziierten Rollen ist hier aufgrund des noch höheren Anspruches noch entscheidender als bei den Varianten a) und b), da die Rolle der "Datenkümmerer" wird hier effektiv um eine nationale Koordinationsaufgabe mit unmittelbarer und tiefer klinischer Relevanz ergänzt.

Bzgl. der Organisation und Umsetzung aller Datenbeschreibungs- und -erfassungsvarianten könnte die Universitätsmedizin, ggf. mit einer Kerngruppe von Pilotstandorten, und gemeinsam mit ausgewählten Partnern anderer Versorgungsstufen in enger Abstimmung mit den relevanten medizinisch-wissenschaftlichen Fachgesellschaften die Funktion einer innovativen Speerspitze übernehmen. Als solche könnte sie die zu schaffenden Standards agil unter verpflichtender Beachtung insbesondere auch des Feedbacks klinischer Anwender aller Versorgungsstufen bis zur breiten Rollout-Reife entwickeln und abstimmen. Wichtig wäre hierbei, von Anfang an auf *eine* gemeinsame, allerdings agil zu adaptierende und mit den versorgungsrelevanten Governance-Strukturen tief integrierte Struktur zu setzen, um den Erfahrungen u.a. aus der MII Rechnung zu tragen, dass die Organisation einer Konvergenz zunächst kompetitiv und divergent organisierter Strukturen aufwändig und langwierig sein kann. Während die Kernaufwände für Koordination, Organisation und Umsetzung der Spezifikationsaktivitäten jedenfalls explizit gegenfinanziert und die erforderliche Strukturanpassung zentral auf

den Weg gebracht werden müsste, könnte die Durchsetzung der Umsetzung in die Breite (bzgl. der Schaffung technischer Lösungen auf der einen Seite und des Rollouts dieser Lösungen auf der anderen Seite) auf zweierlei Weise erfolgen: entweder auf der Basis von Ausschreibungen (eher geeignet, wenn wenige Standorte beteiligt werden sollen) oder durch Erlösincentivierungen, die dann wesentlich breiter wirken würden, wobei auch hier selektivere Mechanismen z.B. über die Aufnahme in die Strukturanforderungen für Zentrumszuschläge denkbar wären.

Um einen unmittelbar versorgungsbezogenen Mehrwert der zu schaffenden Strukturen zu realisieren, könnte ergänzend zu sämtlichen Varianten die Unterstützung des Aufbaus systematischer, systemrelevanter Vigilanz- und Qualitätsmanagementverfahren auf Grundlage der geschaffenen Strukturen bei gleichzeitigem Einsatz der neuen Verfahren zur administrativen Vereinfachung der intensivmedizinischen Erlössicherung erwogen werden:

- Die harmonisierte Datenerfassung und -bereitstellung wird direkt zur Organisation des leistungsbezogenen Abrechnungsprozesses für die Intensivmedizin genutzt, das klinische Personal damit von den aktuell immer weiter zunehmenden, klinisch sinnlosen Dokumentationsverpflichtungen entlastet.

- Es werden Strukturen und Prozesse zur regelmäßigen Kalibrierung (ggf. im Wettbewerb zwischen mehreren Akteuren) prädiktiver Modelle für patientinnen- und patientenzentrierte Outcomes geschaffen.

- Diese Modelle werden zum risikoadjustierten Qualitätsvergleich unter Beteiligung der hierfür relevanten Bundesinstitute eingesetzt. Die resultierende, durch die Fachgemeinschaft zu leistende Analyse der Erkenntnisse liefert einerseits wichtiges Input für lokale und globale Qualitätsverbesserungsmaßnahmen (wenn Standortunterschiede in der gemessenen risikoadjustierten Qualität auf differente Vorgehensweisen zurückgeführt werden können), andererseits aber auch für sowohl die iterative Weiterentwicklung der Dokumentationsstandards (wenn klinisch relevante, aber bisher nicht erfasste Einflussgrößen identifiziert werden) als auch für die Modellbildung (wenn die Effekte bereits erfasster Einflussgrößen nicht adäquat repräsentiert waren).

Der maximale Mehrwert eines solchen Vorgehens wird allerdings erst in Kombination mit der Variante c) (“Umfangreicher Datensatz”) geschöpft.

bb) Bereitstellung interoperabler Standard-Datenelement für die Sekundärnutzung

Die Bereitstellung könnte unter der Voraussetzung geeigneter Spezifikationsprozesse (siehe vorhergehender Abschnitt), die die Regelungsinhalte generieren, durch alle Hersteller vorgeschrieben werden. Dabei könnten bestehende Strukturen und Elemente (beispielsweise Interop Council, ISiK, Medizinische Informationsobjekte) schrittweise zur Umsetzung genutzt werden, und zwar betreffend Frontend (z.B. Eingabemöglichkeiten), Integrationsoptionen (z.B. für Medizin- und Labortechnik) und Backend (z.B. Datenausleitung, Schnittstellen für Entscheidungsunterstützungssysteme und Echtzeitanalytik). Nutzer aller in Deutschland zugelassenen Systeme hätten so einen durch die Fachgemeinschaft getriebenen, ständig aktualisierten Dokumentationsstandard zur Verfügung, der, natürlich, falls dies notwendig sein sollte, unverändert lokal ergänzt werden kann. Idealerweise würden Änderungsanforderungen

allerdings mindestens ergänzend systematisch in den nationalen Entwicklungsprozess eingespeist. Eine systematische Beobachtung der Dokumentation qualitativer klinischer Untersuchungsergebnisse und Beobachtungen mittels Freitexten, die ohnehin immer möglich sein muss, um der Variabilität und Heterogenität des Geschehens in der Patientinnen- und Patientenversorgung Rechnung zu tragen, kann wichtige Informationen über die bereits erreichte Praktikabilität der Dokumentationsstandards liefern und sinnvolle Richtungen für den Weiterentwicklungsprozess identifizieren helfen.

cc) Bereitstellung der Daten unter Beachtung des Datenschutzes

Die Bereitstellung der Daten kann grundsätzlich unter Nutzung aller in Abschnitt C.III.1 beschriebenen Umsetzungsszenarien erfolgen, die dann jeweils mit den dort beschriebenen Chancen und Risiken behaftet sind. Für die Bewertung verschiedener Umsetzungsszenarien wird in diesem Abschnitt eine Auswahl besonders zielführend erscheinender Ansätze betrachtet. Der einwilligungsbasierte Ansatz (Szenario 4) wird im Folgenden nicht weiter betrachtet, da dieser für die Bereitstellung von intensivmedizinischen Daten nur wirksam funktioniert, wenn das Konzept einer bundesweiten "Bürgerinnen- und Bürgerdatenspende" umgesetzt wird, was weit über den Umfang des Aufbaus einer "deutschen MIMIC" hinausgeht. Wir betrachten außerdem den Ansatz einer Anonymisierung allein auf Datenebene (Szenario 1) nicht weiter, da dieser grundsätzlich mit Schwierigkeiten bei der Bewertung der Rechtssicherheit (siehe auch Abschnitt V) behaftet ist und für komplexe Datenarten wie OMICS-Daten, die aus unserer Sicht zwingend Teil einer "deutschen MIMIC" sein sollten, auch nicht wirksam funktionieren kann.

Die Berücksichtigung von hochdimensionalen, faktisch nicht vollständig anonymisierbaren Datenarten wie OMICS-Daten, schon in der grundlegenden Konzeption einer “deutschen MIMIC” erscheint aufgrund der notwendigen Zukunftssicherheit der Planung zwingend notwendig. Es ist sehr wahrscheinlich, dass für eine weitere Verringerung der immer noch sehr hohen intensivmedizinischen Mortalität, und vermutlich der Verbesserung auch weiterer relevanter Zielgrößen der Behandlung, die Nutzung moderner Verfahren wie z.B. genetischer Analysen zur besseren Berücksichtigung individueller Patientinnen- und Patienteneigenschaften eine wichtige Rolle spielen werden[113]. Für diverse klinische Anwendungsbereich mit klinischer Überlappung mit der Intensivmedizin, wie z.B. die Kardiologie, gibt es bereits heute wachsende Evidenz, dass eine Berücksichtigung des individuellen genetischen Profils zur besseren Behandlungsergebnissen führen kann. Eine Infrastruktur für die Ermöglichung translationaler Forschung und datengetriebener Innovation zu schaffen, die ausgerechnet einen zentralen Bereich absehbarer Innovation bereits strukturell ausschließt, erscheint wenig sinnhaft. Gleichzeitig ist bei fehlender Berücksichtigung nicht sicher anonymisierbarer Daten bereits in der frühen Konzeptionsphase eine spätere Nachbesserung kaum mehr möglich, da gerade diese Eigenschaft solcher Daten fundamental andere Strukturanforderungen induziert.

Es verbleiben zwei Szenarien, die unterschiedliche Enden eines Spektrums von stark verteilter Datenhaltung bis zentraler Datenhaltung abbilden: (1) Bereitstellung durch Kombination von Maßnahmen auf Datenebene und Prozessebene, indem – soweit wie möglich anonymisierte – Daten in einer sicheren Verarbeitungsumgebung prozessiert werden (Szenario 2),

[113] See KC. Personalizing Care for Critically Ill Adults Using Omics: A Concise Review of Potential Clinical Applications. Cells. 8. Februar 2023;12(4):541.

sowie (2) Bereitstellung über eine Föderierte Infrastruktur (Szenario 3). Diese werden im Folgenden näher betrachtet. Der erforderliche Ressourceneinsatz für die Bereitstellung ist von der konkreten Ausprägung abhängig.

b) Die Umsetzungsvorschläge im Vergleich

Der Ansatz, der Maßnahmen auf Datenebene mit der Bereitstellung in einer sicheren Verarbeitungsumgebung als Maßnahme auf Prozessebene kombiniert, beinhaltet die Einrichtung und den Betrieb einer sicheren Verarbeitungsumgebung bei einer Datenzugangsstelle. Diese Umgebung erfordert potenziell erhebliche Investitionen in die Infrastruktur, den Betrieb und die Weiterentwicklung. Möglicherweise lassen sich Synergien mit anderen im Aufbau befindlichen Strukturen nutzen, wie bspw. der im Aufbau befindlichen Verarbeitungsumgebung am FDZ oder kommenden EHDS-Strukturen. Ein Schlüsselaspekt dieses Ansatzes ist die Abstimmung der Datenschutzverfahren (siehe Abschnitte C.IV sowie C.V), die sowohl komplex als auch ressourcenintensiv sein kann.

Bei einer dezentralen Bereitstellung durch föderiertes Lernen oder Analysieren, der das genau andere Ende des Lösungsspektrums darstellt, wird auf Federated Learning und Analytics-Methoden zurückgegriffen. Die zentralen Aufwände sind im Vergleich zur ersten Variante moderater, aber es bestehen größere Herausforderungen bei der Anbindung von Institutionen an die erforderlichen Infrastrukturen. Es entsteht auch ein verstärkter Bedarf an Ressourcen auf Seiten der teilnehmenden Institutionen, der potenziell größer sein könnte als der für eine sichere Verarbeitungsumgebung benötigte. Ggf. steigt der Bedarf anders als bei zentralen Verarbeitungsumgebungen mit der Nutzungsintensität auch unabhängig von etwaigen inhaltlich/wissenschaftlich erforderlichen Interaktionen mit den datengenerierenden Institutionen. Auch in diesem Szenario besteht die Möglichkeit, Synergien mit bestehenden oder im Aufbau befindlichen Strukturen zu nutzen. Dazu gehört die föderierte Struktur der Datenintegrationszentren in der MII oder Strukturen wie sie u.a. für das German Human Genome-Phenome Archive (GHGA) aufgebaut werden.

Viele Aufwände skalieren in beiden Szenarien mit der Anzahl der beteiligten, datenliefernden Institutionen. Grundsätzlich gilt hier, dass die Datenbasis umso interessanter wird, je größer diese Anzahl ist. Es sollten Einrichtungen verschiedener Versorgungsstufen, von der Regelversorgung über die Zentralversorgung bis zur Maximalversorgung beteiligt werden.

Was die Aufwände für den Aufbau der Sekundärnutzungsinfrastruktur angeht, sind diese hingegen relativ konstant und wachsen hauptsächlich mit der Nutzungsintensität. Wesentliche Aspekte umfassen die notwendige Hardware sowie die Größe begleitender Governance-Prozesse und Betriebsaufwände. Im föderierten Szenario hängen die Aufwände von der Nutzungsintensität an den jeweiligen Standorten ab, während sie bei zentralen Lösungen vor allem durch die Datenmenge und Integrationsaufwände bestimmt werden. Der wesentliche Unterschied zwischen den beiden Ansätzen liegt in der Zentralisierung versus Dezentralisierung der Datenverarbeitung. Während sichere Verarbeitungsumgebungen hohe initiale und laufende Kosten für die Infrastruktur und den Datenschutz erfordern, bietet die dezentrale Bereitstellung durch föderiertes Lernen Flexibilität und potenziell niedrigere Kosten bei der Anbindung neuer Teilnehmer. Jedoch sind die Aufwände für die Implementierung und Wartung der verteilten Systeme nicht zu unterschätzen. Beide Ansätze nehmen Einschränkungen bei der

Nutzbarkeit der Daten in Kauf, um einen wirksamen Schutz der Anonymität zu gewährleisten (siehe Kapitel V).

Abhängig von Umfang und Sensibilität der verarbeiteten Daten erscheint jedenfalls eine Organisation des Datenzuganges, die datenbezogene und prozessuale Schutzmaßnahmen kombiniert, sinnvoll. Diese sollte unter optimaler Nutzung von Synergien mit bereits bestehenden oder geplanten Strukturen geplant werden.

Im Sinne der international kompetitiven und zukunftssicheren Aufstellung unseres Gesundheitssystems im Allgemeinen und der deutschen Intensivmedizin und intensivmedizinischen Forschung im Besonderen erscheint eine Investition in möglichst weitreichende Anstrengungen zur Datenharmonisierung und -standardisierung, idealerweise im Sinne der oben beschriebenen Variante c), “umfangreicherer Datensatz”, sinnvoll. Eine durch systematische Strukturverbesserung auch messbare Refinanzierung der investierten Aufwände wird dann durch erste Nutzungen der optimierten klinischen Dokumentationsverfahren und damit der Umsetzung von mindestens Teilaspekten dieser Variante möglich. Ob solche tiefgreifenden Innovationen angestrebt werden sollten, ist aber primär eine politische und nicht eine fachlich beantwortbare Frage - auch, weil ein solches Vorgehensmodell tatsächlich auch im internationalen Vergleich hochinnovativ wäre. Auch wenn Teilaspekte des Vorgehensmodells in zahlreichen Varianten in unterschiedlichen Domänen bereits erfolgreich umgesetzt wurden (NICE ICU-Register und Qualitätsmessung, zahlreiche Register, erste leistungsorientierte Vergütungsansätze, etc.), gibt es international keine Präzedenz für ein entsprechend orchestriertes und langfristig gedachtes strategisches Vorgehen. Es wäre insofern zur Risikomitigierung und Maximierung der Erfolgswahrscheinlichkeit dringend angeraten, ein inkrementelles Vorgehen umzusetzen. Dazu könnten notwendige organisatorische und technische Strukturentwicklungen projektartig gefördert und mit monetären Anreizen auf extern messbare Zielgrößen wie (initial) Datenvollständigkeit, im Verlauf dann Indikatoren der Prozessqualität und schließlich Indikatoren der risikoadjustierten Ergebnisqualität kombiniert werden. Geeignete Grundstrukturen, auf denen die erforderliche Governance, Kommunikationsstruktur und technische Infrastruktur aufgebaut werden könnten, finden sich insbesondere im Kontext der MII und des NUM, deren Prozesse für eine nachhaltige Wirksamkeit und Sicherstellung einer gesamtsystemischen Skalierbarkeit aber deutlich enger mit den unmittelbar versorgungsbezogenen Prozessen des Interop Council verzahnt werden müssten, um eine dauerhafte, effiziente und verlässliche Abstimmung zu erreichen.

IV. Fragen zur Anonymität auf Datenebene

Ist es technisch möglich, klinische Datensätze mit Personenbezügen derart aufzubereiten, dass sie rechtlich als anonyme Daten gelten, ohne sie zu verfälschen (im Sinne der Synthetisierung) oder zu aggregieren?

Grundsätzlich ist festzuhalten, dass Re-Identifizierungsrisiken mit keiner Datenschutztechnologie auf Null gesenkt werden können. Die Anonymität von Daten muss deshalb immer nach einem im Ausgangspunkt flexiblen und situationsbezogenen Beurteilungsmaßstab erfolgen, damit klinische Datensätze - auch solche die mittels Anonymisierungs- oder Synthetisierungsverfahren geschützt wurden - als rechtlich anonym gelten können. Legt man einen solchen

Bewertungsmaßstab zugrunde, gibt es durchaus technische Verfahren, die dies ohne Verfälschung erreichen können, auch wenn dadurch potenziell der Informationsgehalt signifikant reduziert wird.

Einerseits gibt es eine ganze Reihe von Anonymisierungs- und Transformationsverfahren, die "wahrheitserhaltend" sind, also keine Daten verfälschen. Wichtige Verfahren in dieser Kategorie umfassen die Kategorisierung, Vergröberung oder Löschung von Informationen. Diese Verfahren müssen nicht unbedingt mit dem Ziel eingesetzt werden, Daten zu aggregieren, sondern können für sich genommen einen gewissen Schutz bieten, der abhängig von der rechtlichen Bewertung auch die Anonymität gewährleisten kann. Dies ist insbesondere dann der Fall, wenn neben reinen Dateneigenschaften auch Prozesseigenschaften oder weitere technische und organisatorische Maßnahmen bei der Einschätzung berücksichtigt werden können (siehe auch Abschnitt V). Andererseits gibt es Anonymisierungsverfahren die Daten zwar verfälschen, nicht aber im Sinne der Synthetisierung. Dazu gehören beispielsweise Anonymisierungsverfahren, die die Differential Privacy-Eigenschaft erfüllen oder Transformationsmethoden wie das „Swapping", bei denen einzelne Werte einer Variable getauscht werden.

Zu guter Letzt können mittels kryptographischer Verfahren aus dem Bereich des Secure-Multiparty-Computing oder der „Homomorphic Encryption" verarbeitete Daten möglicherweise anonym sein. Bei dieser Gruppe von Verfahren werden Daten verschlüsselt und in verschlüsselter Form verarbeitet. Anschließend kann das verschlüsselte Ergebnis in das echte Ergebnis entschlüsselt werden. Es werden also keine Daten offengelegt und es kann dennoch ohne jegliche Verfälschung ein Ergebnis berechnet werden. Für einen Praxiseinsatz zum Betrieb einer mit MIMIC vergleichbaren Struktur sind heutige Verfahren aus diesem Bereich aber nicht skalierbar und flexibel genug.

Mit welchen Ansätzen ließe sich bestimmen, ob die Anonymität von vormals personenbezogenen Daten erreicht wurde?

Grundsätzlich wurden in der Literatur eine Vielzahl von Modellen für die Messung des Grads des Personenbezugs entwickelt (siehe auch Abschnitt C.III.1), von denen viele auch in der Praxis eingesetzt werden. Dazu kann beispielsweise die Eindeutigkeit von als besonders identifizierend geltenden Merkmalskombinationen gehören, das Verhältnis solcher Merkmalskombinationen zu den Eigenschaften der Gesamtpopulation oder auch der Epsilon-Parameter einer Differential Privacy-Implementierung. Dabei ist aber zu beachten, dass fast immer Annahmen über die Natur und Art möglicher Re-Identifizierungsversuche getroffen werden müssen und dass die Privatheit von Patientinnen und Patienten auf unterschiedliche Art und Weise verletzt werden kann.

Die sogenannte Identity Disclosure bezeichnet einen Fall, in dem Patientinnen und Patienten direkt aus den Daten identifiziert werden können. Dies kann durch die Kombination verschiedener, scheinbar nicht identifizierender Informationen geschehen, wie Alter, Geschlecht, seltene Krankheiten, aber auch Merkmalen in Bildern, Signalen oder OMICS meist gefolgt von einem Abgleich mit anderen Datenquellen[114]. Die Folgen einer solchen Enthüllung sind besonders gravierend, da sie nicht nur Verletzungen der informationellen Selbstbestimmung,

[114] Sweeney L. K-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst. 2002;10(5):557-570.

sondern auch Stigmatisierung und Diskriminierung nach sich ziehen können[115]. Wenn man Annahmen über solche Merkmale trifft und diese nicht zu hochdimensional sind, kann das Risiko für solche De-anonymisierungsversuche wirksam gemessen werden.

Die sogenannte Attribute Disclosure bezieht sich auf die Möglichkeit, aus den Daten sensible Attribute einer Person abzuleiten[116]. In einem medizinischen Kontext können diese Attribute Details über den Gesundheitszustand, Diagnosen, Behandlungen, finanzielle Verhältnisse oder persönliche Gewohnheiten umfassen. Die Risiken entstehen oft durch den Zusammenhang zwischen verschiedenen medizinischen Datenpunkten. Zum Beispiel könnte von einer Kombination von Diagnosen, Behandlungsdaten und Medikationsplänen auf spezifische Gesundheitszustände oder Vorlieben geschlossen werden. Solche Informationen könnten von Versicherungen oder Arbeitgebern missbraucht werden, um Entscheidungen zu treffen, die die betroffene Person benachteiligen[117]. Auch für die Messung der Erfolgswahrscheinlichkeiten solcher De-anonymisierungsversuche gibt es mathematische Modelle, die aber ebenfalls Annahmen benötigen. Außerdem ist es herausfordernd, Attribute Disclosure gegen einen - gewünschten - Wissensgewinn abzugrenzen.

Zu guter Letzt bezieht sich die sogenannte Membership Disclosure auf das Offenlegen der Zugehörigkeit einer Person zu bestimmten Gruppen[118]. In einem medizinischen Datensatz könnte beispielsweise die Zugehörigkeit zu einer bestimmten Patientinnen- und Patientengruppe, wie HIV-Positive, aber auch zum Gesamtdatensatz selbst sensible Informationen preisgeben[119]. Um dieses Risiko sinnvoll bestimmen zu können, sind Informationen über die Gesamtpopulation und ebenfalls Annahmen möglich.

Im Kontext der EU-Gesetzgebung wurden unter anderem die Bedrohungen "Aussondern", "Verknüpfen" und "Inferenz" genauer betrachtet, die in enger Verbindung zu den beschriebenen Disclosure Szenarien stehen[120]. "Aussondern" beschreibt die Möglichkeit, Daten zu einer einzelnen Person innerhalb eines Datensatzes zu isolieren. Diese Fähigkeit, Individuen zu differenzieren, ist eine notwendige Voraussetzung für die Identity Disclosure, bei der die Identität einer Person offenbart wird, führt jedoch nicht zwangsläufig dazu. "Verknüpfen" bezieht sich auf den Prozess, bei dem Informationen aus verschiedenen Quellen verknüpft werden können. Dies ist ein zentraler Mechanismus zur Re-Identifizierung von Daten, auch Linkage-Angriffe genannt. Dabei werden anonymisierte Daten mit anderen öffentlich verfügbaren Daten verknüpft, um Identitäten aufzudecken. Insbesondere in hochdimensionalen Daten, die eine Vielzahl von Attributen enthalten, erhöht sich das Risiko solcher Angriffe deutlich[121]. "Inferenz" hingegen bezeichnet die Fähigkeit, durch Analyse und Kombination von vorhandenen

[115] Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Rev. 2010;57:1701.
[116] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data (TKDD). 2007;1(1):3.
[117] Bélisle-Pipon JC, Vayena E, Green RC, Cohen IG. Genetic testing, insurance discrimination and medical research: what the United States can learn from peer countries. Nat Med. 2019;25(8):1198-1204.
[118] Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM.
[119] Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. Data Knowl Eng. 2005;55(3):301-322.
[120] Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. Brussels: European Commission; 2014 Apr.
[121] Willenborg L, de Waal T. Statistical disclosure control in practice. Springer Science & Business Media; 2001.

Daten neue, nicht offensichtliche Informationen über eine Person abzuleiten. Dies kann sowohl zu Attribute Disclosure als auch zu Membership Disclosure führen, je nachdem, ob spezifische Eigenschaften oder die Zugehörigkeit zu einer bestimmten Gruppe aufgedeckt werden. Auch für diese spezifischen Bedrohungen wurden formale Modelle entwickelt, die die zugehörigen Erfolgswahrscheinlichkeiten schätzen[122].

Abschließend sei darauf hingewiesen, dass die beschriebenen - und weitere - Ansätze in anderen Ländern oder Regionen auch praktisch zum Einsatz kommen. In Deutschland verfolgen viele Aufsichtsbehörden hingegen sehr restriktive Sichtweisen auf Anonymität und Ansätze, die (a) Restrisiken zulassen und (b) Annahmen erfordern, konnten sich bisher in der breiten Praxis der medizinischen Forschung nicht durchsetzen. Überdies fehlt es an verlässlichen Vorgaben für Risikogrenzwerte. Ein international gängiger Grenzwert für die maximale Wahrscheinlichkeit, mit der eine Re-Identifizierung einer spezifischen Person bei Versagen aller anderen Schutzmaßnahmen erfolgreich sein darf, ist 9%, was beispielsweise in der External Guidance on the Implementation of Policy 0070 der Europäische Arzneimittel-Agentur genannt ist[123].

V. Mögliche Definitionsansätze für Anonymität

Wie können mögliche Definitionsansätze für „anonymisierte" bzw. „nicht reidentifizierbare Daten" aussehen, die im Spannungsfeld zwischen Datennutzung und Privatsphäre an Forschenden hohe, aber erreichbare und bestimmbare Anforderungen stellen?

EG 26 der DS-GVO definiert „anonyme Informationen" als Informationen, die sich „nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann." Mit dieser Definition einhergeht dann das gängige Verständnis von Anonymität als eine **Eigenschaft von Daten**, abgestellt wird allein auf den Bedeutungsgehalt der Daten als solche, ohne aber das "Gesamtpaket" der jeweiligen Datenverarbeitung in den Blick zu nehmen. Eine solche isolierte Betrachtungsweise stellt aber eine große Herausforderung dar und ist als rechtlicher Mechanismus für die Nutzung von Daten für Forschungszwecke kaum operationalisierbar.

Die Grundproblematik liegt darin, dass Daten auf Individualebene aufgrund ihrer inhärent eindeutigen Merkmalskombinationen oder Zusammenhänge stets eine potenzielle Zuordnung zu Personen erlauben. Daten, die Informationen über Personen oder Gruppen von Personen beschreiben, können niemals absolut anonym sein, da dies einen Netto-Informationsgehalt oder einen Wahrheitsgehalt von Null implizieren würde. Dies liegt daran, dass jegliche Informationen zu Personen oder Personengruppen, die eine Form von Wert oder Nutzen bieten, potenziell auch konkreten Personen zugeordnet werden können oder Informationen über diese offenlegen, gerade auch in Verbindung mit Daten aus anderen Quellen. In der Praxis bedeutet dies, dass die vollständige Entfernung oder Verzerrung aller theoretisch möglichen

[122] Cohen A, Nissim K. Towards formalizing the GDPR's notion of singling out. Proc Natl Acad Sci U S A. 2020;117(15):8344-8352.
[123] External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. 20 September 2017. EMA/90915/2016. Version 1.3 Veröffentlicht: 22. September 2017

Personenbezüge aus einem Datensatz diesen in seiner Gesamtheit nutzlos macht. Die Anonymisierung von Daten ist insofern stets ein Prozess, der Re-Identifizierungsrisiken gegen den Informations- oder Wahrheitsgehalt abwägen muss [124,125]. Damit ist dann aber auch die von den Aufsichtsbehörden noch immer praktizierte Gleichstellung von "Anonymität" und "Re-Identifizierungsrisiko identisch Null" weder praktikabel noch operationalisierbar, da Daten aufgrund ihres Informationsgehalts stets ein unvermeidliches Rest-Re-Identifizierungsrisiko in sich tragen.

Eine operationalisierbare Legaldefinition könnte in der Auffassung von Anonymität als einer Eigenschaft nicht von Daten, sondern von Verarbeitungstätigkeiten bzw. Prozessen liegen. An einem Prozess sind neben Daten auch diverse Akteure und Systeme beteiligt und sie generieren ein bestimmtes Ergebnis. Bei der Bewertung der Frage, ob ein Prozess die Anonymität der Betroffenen wahrt, könnten neben den Dateneigenschaften natürlicherweise Aspekte wie die Vertrauenswürdigkeit der beteiligten Akteure sowie die auf verschiedenen Ebenen ergriffenen technischen und organisatorischen Maßnahmen berücksichtigt werden. Die Eigenschaften der Daten spielen zwar eine Rolle, sind jedoch nicht allein ausschlaggebend für die Gewährleistung von Anonymität. Des Weiteren umfasst ein Prozess stets die Aktionen aller beteiligten Akteure und bietet somit potenziell die Möglichkeit, stärker zu berücksichtigen, welche Mittel durch Prozessbeteiligte nach aller Wahrscheinlichkeit eingesetzt werden.

Eine solche "prozessorientierte Sicht" auf den Anonymitätsbegriff ist auch mit den Grundwertungen der DS-GVO vereinbar. Nach deren EG 26 sollen für die Feststellung, ob eine natürliche Person identifizierbar ist, alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person "**nach allgemeinem Ermessen wahrscheinlich genutzt werden**", um eine Person direkt oder indirekt zu identifizieren. Ausschlaggebend ist also die Identifizierbarkeit im konkreten Datenverarbeitungsprozess. Entscheidend ist, ob es unter den jeweiligen Rahmenbedingungen einer Datenverarbeitung mit einem realistischen Aufwand an Zeit, Kosten und Arbeitskraft möglich ist, Informationen einer bestimmten Person zuzuordnen. Erfordert die Identifizierung dagegen einen unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft oder ist es von vornherein unwahrscheinlich, dass der Versuch einer Identifizierung überhaupt von irgendjemandem unternommen wird, ist eine Identifizierbarkeit abzulehnen.[126] Das Risiko einer Identifizierung ist dann — in den Worten des EuGH — „de facto vernachlässigbar".[127] Die rein hypothetische Möglichkeit, eine Person zu bestimmen, reicht gerade nicht aus, um diese Person als identifizierbar und damit Daten als personenbezogen einzuordnen.[128]

---

[124] Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Review, 57, 2010.
[125] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211-407.
[126] So auch Positionspapier des BfDI zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche (Stand: 29.6.2020), S. 4.
[127] EuGH v. 19.10.2016, DuD 2017, 42, 44 – Breyer.
[128] In diesem Sinne schon Art.-29-Datenschutzgruppe, Personenbezogene Daten, WP 136 (2007), S. 17.

Ein prozessorientiertes Verständnis von Anonymität spiegelt zudem auch die Realität der Datenverarbeitung in der Forschung wider, in der der Kontext und die Absicht hinter der Datennutzung oftmals entscheidender sind als dateninhärente Merkmale. Das Hauptinteresse der Forschenden liegt im Regelfall nicht bei den Daten selbst, sondern vielmehr in der Durchführung von Verarbeitungsprozessen, die zu, bspw. statistischen, Ergebnissen führen und einen Erkenntnisgewinn ermöglichen. Überdies wird die umfassendere Sicht auf den Anonymitätsbegriff auch der Komplexität und Dynamik moderner Datenverarbeitungen besser gerecht.

Auch das auf verschiedenen Ebenen und inkl. auch der Prozessebene angesiedelte und weit verbreitete Five Safes Framework[129] für das sichere Teilen von Daten, sowie aktuelle Entwicklungen im Bereich der Privacy-Enhancing Technologies liefern Indizien dafür, dass eine prozessorientierte Betrachtung von Anonymität Vorteile bietet. Ein Beispiel hierfür ist das Konzept der Differential Privacy (siehe Abschnitt C.III.1.a.ff), die in vielen Teilen der Wissenschaft aufgrund des starken Schutzes der Privatheit als "Goldstandard" gilt. Das Kernprinzip von Differential Privacy basiert dabei darauf, Anonymität als eine Eigenschaft von auf Daten ausgeführten mathematischen Funktionen, also Verarbeitungstätigkeiten, und nicht als Eigenschaft der Daten selbst zu definieren. Im Gegensatz zu traditionellen Anonymisierungsverfahren ist Differential Privacy deutlich robuster. Ein weiteres Beispiel ist das Konzept der homomorphen Verschlüsselung, bei der mittels spezieller Prozesse Daten in verschlüsselter Form verarbeitet werden.

Schließlich lässt sich der spezifisch situations- bzw. prozessbezogene Beurteilungsmaßstab für die Personenbeziehbarkeit von Daten auch für den Fall einer Pseudonymisierung von Daten heranziehen, die insbesondere dann von Relevanz ist, wenn für eine MIMIC (auch) auf das Konzept der Treuhandstellen zurückgegriffen werden soll (siehe das beschriebene Szenario 2). Zwar gelten nach EG 26 der DS-GVO pseudonymisierte Daten im Ausgangspunkt weiterhin als personenbezogene Daten. Letztlich ist jedoch auf die Rahmenbedingungen des gesamten Datenverarbeitungsprozesses abzustellen, um bestimmen zu können, im Verhältnis zu welchem Verantwortlichen pseudonymisierte Daten als personenbezogene Daten einzuordnen sind. Ist es einem Verantwortlichen, an den pseudonymisierte Daten weitergegeben werden, praktisch unmöglich, mit einem verhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskräften an die Zuordnungsregel zu gelangen, und gibt es für diesen auch keinen rechtlich gangbaren Weg, die Zuordnungsregel zu erfahren, so weisen für diesen Verantwortlichen die pseudonymisierten Daten keinen Personenbezug auf.[130] Zu berücksichtigen ist insoweit vor allem auch, ob es sich bei der Stelle, die die Zuordnungsregel für ein Pseudonym vergeben hat und dieses Pseudonym verwaltet, um eine Stelle handelt, die mit besonderen Vertraulich-

[129] Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. Economics Working Paper Series. 2016;1601:28.

[130] So schon Roßnagel A. Pseudonymisierung personenbezogener Daten. ZD 2018, 243 und jüngst auch EuG v. 26.04.2023 – T-557/20, BeckRS 2023, 8240. S. in diesem Sinne auch schon die Breyer-Entscheidung des EuGH v. 19.10.2016 - C-582/14, BeckRS 2016, 82520, wonach es für die Frage einer Personenbeziehbarkeit von Daten darauf ankommt, ob die verantwortliche Stelle über **rechtliche Mittel** verfügt, die es ihr ermöglichen, auf Zusatzinformationen zurückzugreifen, über die eine andere Stellen verfügt, um anhand dieser Zusatzinformationen die betreffende Person dann identifizieren zu können. Anknüpfend an diese Entscheidung wiederum jüngst auch EuGH v. 9.11.2023 - C-319/22, wonach es sich bei einer Kennnummer um ein personenbezogenes Datum der hinter dieser Nummer stehenden Person handelt, sofern derjenige, der Zugang zu dieser Nummer hat, "**über Mittel verfügen könnte, die es ihm ermöglichen**", die Kennnummer zur Identifizierung der betreffenden Person zu nutzen.

keitspflichten und -rechten ausgestattet ist, und daher mit hinreichender Sicherheit ausgeschlossen werden kann, dass diese Stelle den Personenbezug der pseudonymisierten Daten gegenüber anderen Stellen offenlegt. Es handelt sich bei diesem – relativen – Ansatz um nichts anderes als eine konsequente Fortsetzung der EuGH-Rechtsprechung zur Personenbezogenheit von Daten, beginnend mit der Breyer-Entscheidung aus 2016. Jeweils stellt der EuGH (und mit ihm auch das EuG) darauf ab, ob diejenige Stelle, für die Daten lediglich mit einer IP-Adresse, einer sonstigen Kennnummer oder einem Pseudonym verknüpft sind, ”bei vernünftiger Betrachtung über Mittel verfügt“, die es ihr ermöglichen, diese Kennung letztlich doch auch einer bestimmten Person zuordnen zu können.

Schlussthese: Um zu einer praktisch operationalisierbaren Definition von Anonymität auf Prozessebene zu kommen, müsste zunächst ein Anonymitätsbegriff definiert werden, der sich an EG 26 der DS-GVO orientiert, sich jedoch anstatt auf Daten auf spezifische Verarbeitungsprozesse bezieht. Begleitet werden muss dies durch Festlegung definierter Prüfverfahren für die Feststellung der Wahrung der so definierten Anonymität der Betroffenen im Rahmen entsprechender Verarbeitungsprozesse. Dies könnte ergänzt werden durch Kataloge an vorgeschlagenen Maßnahmen zur Absicherung und Gestaltung solcher Prozesse.

Abkürzungsverzeichnis

- BayKrG: Bayerisches Krankenhausgesetz
- BDSG: Bundesdatenschutzgesetz
- BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte
- BfDI: Bundesbeauftragter für den Datenschutz und die Informationsfreiheit
- BMG: Bundesministeriums für Gesundheit
- BremDSGVOAG: Bremisches Ausführungsgesetz zur EU-Datenschutz-Grundverordnung
- DGAI: Deutschen Gesellschaft für Anästhesiologie und Intensivmedizin
- DIVI: Deutschen Interdisziplinären Vereinigung für Intensiv- und Notfallmedizin
- DSG NRW: Datenschutzgesetz Nordrhein-Westfalen
- DS-GVO: Datenschutz-Grundverordnung
- EG: Erwägungsgrund
- EHDS: European Health Data Space, Europäischer Gesundheitsdatenraum
- EKG: Elektrokardiogramm
- ePA: elektronischen Patientinnen- und Patientenakte
- EuGH: Europäischer Gerichtshof
- FDZ: Forschungsdatenzentrum Gesundheit
- GAN: Generative Adversarial Network
- GDNG: Gesundheitsdatennutzungsgesetz
- GG: Grundgesetz
- GHGA: German Human Genome-Phenome Archive
- HIPAA: Health Insurance Portability and Accountability Act
- InEK: Institut für das Entgeltsystem im Krankenhaus
- IQTiG: Institut für Qualitätssicherung und Transparenz im Gesundheitswesen
- IQWiG: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
- ISiK: Informationstechnische Systeme in Krankenhäusern
- KI: Künstliche Intelligenz
- LSTM: Long Short-Term Memory
- MII: Medizininformatik-Initiative
- MIMIC: Medical Data Mart for Intensive Care
- MRT: Magnetresonanztomographie
- NDSG: Niedersächsisches Datenschutzgesetz
- NLP: Natural Language Processing, Computerlinguistik
- NUM: Netzwerk Universitätsmedizin
- RKI: Robert Koch-Institut
- RNN: Rekurrentes neuronales Netzwerk
- RWD: Real World Data ,“Reale-Welt-Daten”
- SGB V: Sozialgesetzbuch (SGB) Fünftes Buch (V)
- TRE: Trusted Research Environment, Sichere Verarbeitungsumgebung
- USA: Vereinigte Staaten von Amerika

Glossar

- Anonymisierung: Der Prozess der Entfernung oder Veränderung personenbezogener Daten, sodass die Identität der Personen nicht mehr feststellbar ist.
- Anonymisierungsverfahren: Methoden zur Veränderung personenbezogener Daten, um die Identifizierung betroffener Personen zu verhindern.
- Attribute Disclosure: Die Möglichkeit, sensible Attribute einer Person aus einem Datensatz abzuleiten.
- Breite Forschungseinwilligung (Broad Consent): Eine Einwilligungsform, die es ermöglicht, Daten für Forschungszwecke zu nutzen, auch ohne spezifische Zustimmung für jedes einzelne Forschungsvorhaben.
- Datensynthetisierung: Techniken zur Erzeugung künstlicher Daten, die echte Daten ersetzen können.
- Datentreuhandstellen: Unabhängige Stellen, die Identitätsdaten pseudonymisieren, Einwilligungen verwalten oder Datenanalysen im Auftrag durchführen.
- Föderiertes Lernen: Ein Ansatz der Datenverarbeitung, bei dem Algorithmen dezentral auf Daten trainiert werden, ohne dass die Daten ihren ursprünglichen Ort verlassen.
- Generative Adversarial Networks: Ein Ansatz im maschinellen Lernen, bei dem zwei neuronale Netzwerke gegeneinander antreten, um synthetische Daten zu erzeugen, die von realen nicht zu unterscheiden sind.
- Homomorphe Verschlüsselung: Eine Form der Verschlüsselung, die es erlaubt, Berechnungen direkt auf verschlüsselten Daten durchzuführen, ohne sie zu entschlüsseln.
- Identity Disclosure: Die Zuordnung der Identität einer Person zu einem Datensatz.
- Membership Disclosure: Die Offenlegung der Zugehörigkeit einer Person zu einer bestimmten Gruppe oder einem Datensatz.
- MIMIC (Medical Information Mart for Intensive Care): Eine öffentlich zugängliche Datenbank, die detaillierte Daten von Patientinnen und Patienten einer Intensivstation beinhaltet, einschließlich physiologischer Messungen, Laborergebnissen und mehr, um Forschung im Gesundheitswesen zu unterstützen.
- Privacy Tests: Methoden zur Bewertung des Risikos der Re-Identifizierbarkeit, u.a. bei pseudonymisierten, anonymisierten oder synthetischen Daten.
- Re-Identifizierungsrisiko: Das Risiko, dass Daten einer spezifischen Person zugeordnet werden können.
- Sichere Verarbeitungsumgebungen: Geschützte Umgebungen, die sensible Daten vorhalten und nur über speziell gesicherte Mechanismen für die Auswertung zugänglich machen.