Iris Pigeot, Holger Fröhlich, Timm Intemann, Guido Prause, Marvin N. Wright

KI und die Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten (NFDI4HEALTH)

In:
Dössel, Olaf / Schäffter, Tobias / Rutert, Britta (Hrsg.): Künstliche Intelligenz in der Medizin.
ISBN: 978-3-949455-18-6
Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, 2023
S. 62-74
(Denkanstöße aus der Akademie : eine Schriftenreihe der Berlin-Brandenburgischen Akademie der Wissenschaften ; 11)

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-38075

KI UND DIE NATIONALE FORSCHUNGSDATENINFRASTRUKTUR FÜR PERSONENBEZOGENE GESUNDHEITSDATEN (NFDI4HEALTH)

Iris Pigeot, Holger Fröhlich, Timm Intemann, Guido Prause, Marvin N. Wright

Hintergrund

Die Digitalisierung hat auch im Gesundheitswesen zu deutlich wachsenden Datensätzen geführt, die intelligente Lösungen im Forschungsdatenmanagement erfordern und zu deren Auswertung Standardmethoden der Statistik bereits jetzt nicht mehr ausreichen: Echtzeitmessungen und eine hohe Anzahl von Messzeitpunkten führen zu extrem hochdimensionalen Daten. Neue Datenstrukturen wie von Fotos/Videos, Barcode-Scans, GPS-Standorten etc. müssen in der Auswertung berücksichtigt werden. Dabei stellt uns die Verknüpfung verschiedener Datenquellen vor besondere Herausforderungen aufgrund verschiedener Messmethoden, heterogener Datenstrukturen und verschiedener Pseudonyme. Dazu kommen methodische Probleme wie die Akkumulation von Rauschen, Messfehlern, versteckte Einflussfaktoren, nicht-lineare Dynamiken, zweifelhafte Korrelationen und schließlich die schiere Größe des Datensatzes, durch die sich letztendlich (fast) jede untersuchte Fragestellung als – zumindest nominell – statistisch signifikant erweist. Das Problem wird von Taleb (2021) gut auf den Punkt gebracht: „I am not saying here that there is no information in big data. There is plenty of information. The problem – the central issue – is that the needle comes in an increasingly larger haystack."

Unter Anwendung der richtigen Auswertungsinstrumente bieten diese umfangreichen Datensätze allerdings ein enormes Potenzial, die Entstehung von Erkrankungen und die Wirksamkeit von Präventionsmaßnahmen umfassend zu erforschen, z. B. unter Einbeziehung diagnostischer und genetischer Informationen, Lebensstilen und Umweltdaten. Ein weiteres Anwendungsgebiet ist die sich entwickelnde Präzisionsmedizin, durch die künftig eine genauere Diagnostik, ein zunehmend lückenloses Krankheitsmonitoring (etwa über digitale Gesundheitsanwendungen) und bessere individualisierte Behandlungsangebote realisiert werden könnten. Um dies zu ermöglichen, müssen Gesundheitsdaten kuratiert und für die Forschung strukturiert bereitgestellt werden und ihre Qualität muss dokumentiert und gesichert sein. Dabei erfordern personenbezogene Gesundheitsdaten aufgrund ihrer Sensibilität einen hohen Schutz, insbesondere, wenn durch die Verknüpfung verschiedener Datensätze eine hohe Informationstiefe entsteht.

Beispiele für den Einsatz von KI im Gesundheitsbereich

Um die oben skizzierten Probleme in der Auswertung solcher Datensätze bewältigen zu können, werden immer häufiger Methoden der Künstlichen Intelligenz (KI) wie Maschinelles Lernen, Data-Mining-Methoden oder Netzwerk-Analysen eingesetzt. Maschinelle Lernverfahren können z. B. helfen, unerwünschte Arzneimittelwirkungen (UAW) anhand von Abrechnungsdaten gesetzlicher Krankenversicherungen aufzudecken, wobei im Unterschied zu den Spontanmelderegistern kein Verdacht auf eine UAW vorliegen muss. Zudem können dabei Komorbiditäten und Komedikationen adäquat berücksichtigt werden (Foraita et al. 2018). Maschinelle Lernverfahren erlauben zudem eine multivariate Modellierung auch extrem hochdimensionaler Daten, wie sie beispielsweise bei der Untersuchung von Zusammenhängen zwischen genetischen Varianten und Krankheiten unter Berücksichtigung von Gen-Gen- und Gen-Umwelt-Interaktionen entstehen, wobei die Methoden zumeist auf die Interpretation von Zusammenhängen abzielen, nicht auf deren Vorhersage (Watson, Wright 2021; Boulesteix et al. 2020). Maschinelle Lernverfahren lassen sich auch zur Auswertung longitudinaler Daten mit vielen Messzeitpunkten und von Überlebenszeiten mit konkurrierenden Ereignissen einsetzen. So werden anhand von Registerdaten von Statistics Denmark ältere Menschen mit einem hohen Risiko für Pflegebedürftigkeit identifiziert, wodurch eine bessere Steuerung von Pflegeangeboten in Dänemark ermöglicht werden könnte (Wright et al. 2021). Weitere Beispiele für den Einsatz von KI im Gesundheitsbereich umfassen das Clustern von Krankheitsverläufen anhand multivariater Endpunkte (de Jong et al. 2019), Risikomodelle für chronische Erkrankungen unter Berücksichtigung multipler individueller Faktoren (Khanna et al. 2018, Linden et al. 2021) sowie Vorhersagen für das individuelle Ansprechen auf ein bestimmtes Medikament (de Jong et al. 2021).

Data Sharing im Gesundheitsbereich

Die genannten Beispiele unterstreichen nicht nur die Bedeutung von KI bei der Auswertung großer Datensätze im Gesundheitswesen, sondern auch die Bedeutung der Datensätze selbst und ihrer Bereitstellung für statistische Analysen zum Allgemeinwohl einer Bevölkerung. Typischerweise ist eine umfassende Nutzung des Potenzials von Forschungsdaten z. B. im Rahmen eines Forschungsprojekts mit engem Fokus und begrenzter Dauer gar nicht möglich. Mit der Bereitstellung solcher Daten für eine Zweitnutzung ergibt sich damit die Chance zur Untersuchung von zum Zeitpunkt des Projekts nicht absehbaren Forschungsfragen. Außerdem

lassen sich so verschiedene Studien zu einer großen Studie poolen, wodurch insbesondere seltene Erkrankungen, kleine Effekte wie z. B. genetische Risiken oder heterogene Bevölkerungsgruppen untersucht werden können. Auch lässt sich durch die Verknüpfung verschiedener personenbezogener Datensätze (Record Linkage) ein umfassenderes Bild eines bestimmten Krankheitsgeschehens zeichnen.

Im Sinne einer effizienten Ressourcennutzung forderte daher die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) bereits 2007 einen einfachen Zugang zu Forschungsdaten für die gesamte wissenschaftliche Community. 2016 wurden von Wilkinson et al. die sogenannten FAIR-Principles publiziert, wonach Daten auffindbar (**F**indable), zugänglich (**A**ccessible), interoperabel (**I**nteroperable) und wiederverwendbar (**R**eusable) sein sollen. Damit einher gehen Forderungen nach klaren Qualitätskriterien und Standards. Die FAIR-Prinzipien wurden mittlerweile von Förderinstitutionen und Forschungsorganisationen mit dem Leitgedanken „as open as possible, as closed as necessary" übernommen, so dass es nicht verwundert, dass von vielen Seiten Anstrengungen für den Aufbau einer Infrastruktur unternommen werden, die eine Bereitstellung und Zweitnutzung von qualitätsgesicherten Forschungsdaten ermöglicht. In Deutschland wurde auf Empfehlung des Rats für Informationsinfrastrukturen (RfII 2016) und nach einer entsprechenden Bund-Länder-Vereinbarung vom 26. November 2018 (GWK 2018) im Jahr 2020 mit dem Aufbau einer Nationalen Forschungsdateninfrastruktur (NFDI 2021) begonnen. Auf diese Weise soll ein bundesweites, verteiltes und wachsendes Netzwerk zur systematischen Erschließung wissenschaftlicher Datenbestände sowie zur nachhaltigen Sicherung und Erhöhung der Zugänglichkeit dieser Datenbestände entstehen. Damit soll aber nicht nur die nationale Vernetzung vorangetrieben werden, sondern auch eine Vernetzung auf internationalem Niveau. Insgesamt sollen bis zu 30 Konsortien in drei Ausschreibungsrunden mit einem Budget von bis zu 90 Mio. € pro Jahr im Endausbau gefördert werden. Das Direktorat (Leitung: York Sure-Vetter) ist in Karlsruhe angesiedelt.

Als eines der Konsortien, die bereits in der ersten Ausschreibungsrunde gefördert wurden, ist NFDI4Health zum Aufbau einer Nationalen Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten (Leitung: Juliane Fluck, Stellvertr.: Iris Pigeot) bereits im Oktober 2020 mit einer zunächst 5-jährigen Laufzeit an den Start gegangen. NFDI4Health will Lösungen erarbeiten, die den besonderen Herausforderungen von personenbezogenen Gesundheitsdaten Rechnung tragen: Auch wenn es sich in den meisten Fällen bereits um strukturierte und qualitätsgesicherte Daten handelt, die insbesondere in Forschungsprojekten gemäß definierten Erhebungsprotokollen erhoben wurden, ergeben sich besondere

Probleme: Denn es handelt sich dabei in der Regel um „lebende" Datenkörper mit einem Bedarf an ständiger Pflege, Aktualisierung und Fortschreibung. Hinzu kommt, dass diese Daten besonders sensibel und damit besonders schützenswert sind. Zudem ist die Generierung absoluter Anonymität aufgrund der üblicherweise in den Studien erfolgten tiefgehenden Phänotypisierung unmöglich. Die Nutzungsmöglichkeiten der Daten sind darüber hinaus durch die jeweils erteilte informierte Einwilligung der Studienteilnehmer:innen beschränkt. Dazu kommt, dass ihre Auffindbarkeit trotz existierender Portale wie re3data.org oder DataCite beeinträchtigt ist und eine Metadatenbeschreibung häufig fehlt. Die gemäß den FAIR-Prinzipien geforderte Interoperabilität zwischen verschiedenen Datenquellen ist in der Regel nicht gegeben, da jede Institution ihre eigenen Standards anwendet. Auch sind die Möglichkeiten für einen geregelten Datenzugang sehr eingeschränkt, wodurch unter anderem Data-Mining-Ansätze und Maschinelle Lernverfahren routinemäßig nicht angewendet werden können.

NFDI4Health hat sich daher zum Ziel gesetzt, (1) die Auffindbarkeit von Gesundheitsdaten durch den Aufbau eines Central Search Hub zu verbessern, Datenpublikationen zu unterstützen, Metadaten zu standardisieren und so zur Verbesserung der Interoperabilität beizutragen; (2) einen übergeordneten Datenzugangs- und Datennutzungsprozess (Central Data Access Point) zu implementieren; (3) Prozesse aufzusetzen, die eine ausschließliche Nutzung im Einklang mit den gegebenen Einwilligungserklärungen und mit den geltenden Datenschutzrichtlinien gewährleisten; (4) Dienste weiterzuentwickeln, die einen kontrollierten Zugriff auf verteilt vorliegende Daten mittels Analysetools erlauben; und (5) Dienste für eine dynamische und sichere Verknüpfung von Primär-, Sekundär- und Registerdaten zu entwickeln. In all diese Aktivitäten sollen die Nutzer:innen eng eingebunden werden, um so eine große Akzeptanz und Nachhaltigkeit der aufgesetzten Strukturen zu erreichen.

Der Einsatz von KI-Methoden in der NFDI4Health

Zur Erreichung dieser Ziele werden auch vielfach KI-Methoden eingesetzt, wie im Folgenden an drei Beispielen aus sehr unterschiedlichen Bereichen illustriert wird.

KI zur Verarbeitung medizinischer Bilddaten: Wie bereits zu Beginn erwähnt, stehen durch die Digitalisierung neue Datenstrukturen wie z. B. medizinische Bilddaten in Verbindung mit anderen medizinischen Daten für die statistische Analyse zur Verfügung. Dabei zielt Radiomics unter Ausnutzung von KI-Methoden, speziell

von Deep Learning (DL)-Ansätzen, auf die Identifikation quantitativer prädiktiver Bildgebungsparameter (s. Abbildung 1). Plattformbasierte Radiomics-Ansätze weisen eine Reihe von Vorteilen auf, u. a. Datentransparenz, Zuverlässigkeit, die Verwendung von Standards und nicht zuletzt die breite Einbindung der Community (Overhoff et al. 2021).

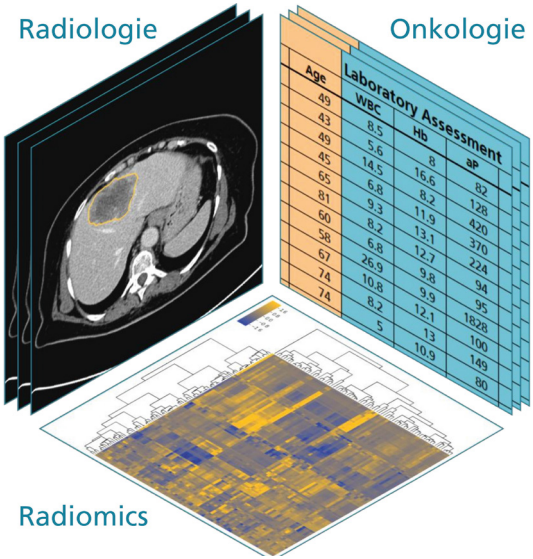

Abb. 1: Radiomics zur Aufdeckung von Zusammenhängen zwischen radiologischen und klinischen Daten

Ziel ist im Rahmen der NFDI4Health, eine KI-basierte Pilot-Radiomics-Plattform für automatisierte und interaktive Analysen sowie für die automatisierte Qualitätssicherung von biomedizinischen Bildern als Service für die Community bereitzustellen. Auf Grundlage eines geeigneten Datenschutz- und Nutzungskonzepts soll diese Plattform zudem Bildanalysen mit Hilfe von DL-Radiomics-Modulen sowie die kollaborative Kuratierung und Annotation der Daten ermöglichen. Dabei erlaubt der Einsatz von KI eine einfache Übertragbarkeit bzw. Anpassung an neue epidemiologische und klinische Daten sowie eine kontinuierliche Optimierung insbesondere durch föderiertes Lernen. Ein Prototyp für eine solche Plattform wurde für die Auswertung von computertomographischen Lungendaten im Zu-

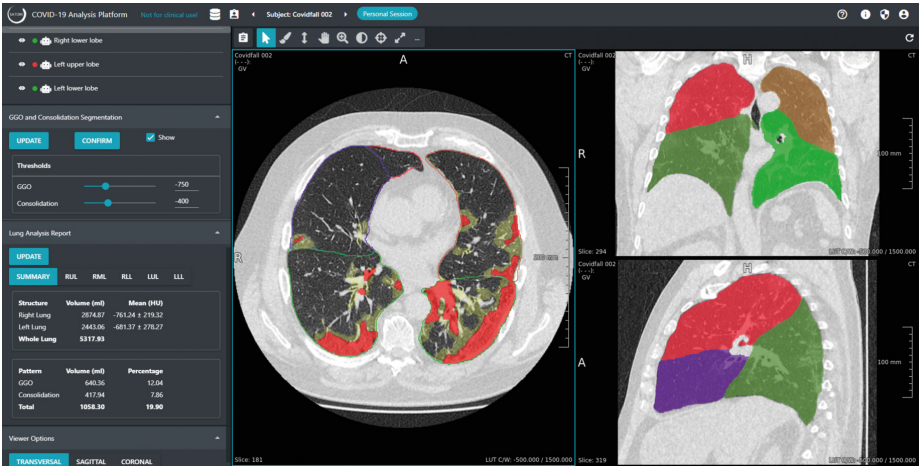

Abb. 2: Prototypische Plattform zur interaktiven Sichtung und Kuratierung von computertomographischen Daten der Lunge

sammenhang mit COVID-19 (Task Force COVID-19; Schmidt et al. 2021, Lessmann et al. 2021) erstellt (s. Abbildung 2).

KI zur Erzeugung synthetischer Daten: Eine Möglichkeit, sensible Gesundheitsdaten einfacher unter Einhaltung der bestehenden Anforderungen des Datenschutzes zu teilen, besteht in der KI-basierten Erzeugung synthetischer Daten, die den realen Daten möglichst „ähnlich" sind, aber unter realistischen Annahmen keinen Rückschluss auf die wahren Individuen erlauben. Dazu wird z. B. mit Hilfe eines modularen Bayes'schen Netzwerkes (VAMBN – Gootjes-Dreesbach et al. 2020, Sood et al. 2020), das gegebenenfalls auch mit Differentialgleichungen zur Beschreibung von Krankheitsdynamiken kombiniert werden kann (MultiNODE – Wendland et al. 2021), eine mathematische Repräsentation der Originaldaten generiert, aus der dann die synthetischen Daten in beliebiger Quantität erzeugt werden können (s. Abbildung 3).

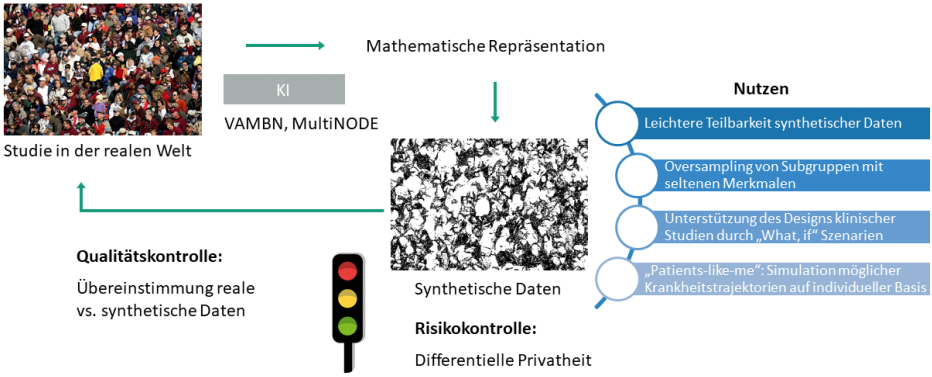

Abb. 3: Erzeugung synthetischer Daten mit Hilfe von KI-Verfahren

Dies geschieht innerhalb der datenhaltenden Organisation. Dabei wird davon ausgegangen, dass externe Nutzer:innen synthetischer Daten keine Kenntnis darüber besitzen, welche realen Individuen in den Originaldaten erfasst sind und dass sie insbesondere auch keinen eigenen Datensatz mit einer großen Zahl personenbezogener Merkmale dieser Personen besitzen, denn in solch einem (äußerst unwahrscheinlichen) Fall wäre unter Umständen eine Re-Identifikation einer zur Erzeugung der synthetischen Daten genutzten Person aus den synthetischen Daten durch Ähnlichkeitsbetrachtung möglich.

Der dargestellte Ansatz ist in den zitierten Publikationen schon erfolgreich angewandt worden, um sehr realistische synthetische Daten aus verschiedenen Parkinson- und Alzheimerstudien zu generieren. Der längerfristige strategische Nutzen ist offensichtlich: (1) Synthetische Daten könnten leichter als reale Daten geteilt werden, da das Risiko der Re-Identifizierbarkeit erheblich reduziert ist. Im Vergleich zu gängigen Anonymisierungsverfahren bieten synthetische Daten dabei den Vorteil, dass wichtige Informationen in den Daten wesentlich besser abgebildet werden, der Nutzen für Forschungszwecke also deutlich höher ist. (2) Subgruppen mit seltenen Merkmalen könnten überrepräsentiert werden, was speziell für die Erforschung seltener Erkrankungen von Vorteil wäre. (3) Die Planung klinischer Studien könnte durch das Durchspielen von „Was wäre, wenn"-Szenarien unterstützt werden. (4) Zudem könnten mögliche Trajektorien der künftigen Krankheitsentwicklung auf individueller Basis simuliert werden, wodurch Patienten besser informiert werden könnten.

Natürlich bedürfen synthetische Daten einer Qualitäts- und Risikokontrolle. Zum einen muss durch einen Vergleich der realen mit den synthetischen Daten innerhalb der datenhaltenden Organisation geprüft werden, ob diese weitestgehend „übereinstimmen", also zu vergleichbaren statistischen Verteilungen und daraus ableitbaren Aussagen führen. Zum anderen muss aber auch das Risiko der Re-Identifizierung im Sinne der differentiellen Privatheit überprüft bzw. kontrolliert werden.

Anwendung von KI in verteilten Datenanalysen: Eine weitere Möglichkeit, den Datenschutz beim Teilen von Daten besser zu gewährleisten, besteht darin, diese selbst nicht weiterzugeben, sondern bei den Dateneignern zu belassen und sie durch geeignete Algorithmen, die auf Maschinellen Lernverfahren basieren, föderiert zu analysieren und so virtuell zusammenzuführen. Die dafür nötige Infrastruktur wird mit dem sogenannten Personal Health Train bereitgestellt. Dieser wird sowohl in dem Fall eingesetzt, bei dem dieselben Variablen für verschiedene Individuen in verschiedenen Datensätzen erfasst wurden, die Datensätze also horizontal verteilt sind, als auch in dem Fall, bei dem verschiedene Variablen für dieselben Individuen in verschiedenen Datensätzen erfasst wurden, die Datensätze also vertikal verteilt sind (Deist et al. 2017, Van Soest et al. 2018). Das konkrete Modell für verteiltes Lernen hängt dabei von dem eingesetzten Algorithmus und der jeweiligen Implementation ab. So kann etwa iterativ vorgegangen werden: Es finden zunächst separate und parallele Anpassungen von Modellparametern für jeden Datensatz statt. Anschließend werden diese durch einen allgemeinen Masterknoten verglichen und angepasst, falls keine Konvergenz erreicht wurde. Der Personal Health Train soll im Rahmen der NFDI4Health erprobt und entsprechend weiterentwickelt werden.

Möglichkeiten des Record Linkage

In Deutschland ist die Verknüpfung von personenbezogenen Sozial- und Gesundheitsdaten auf Basis personenidentifizierender Variablen mit sehr hohen datenschutzrechtlichen Anforderungen und einem hohen administrativen Aufwand verbunden: Zur Verknüpfung von Primärdaten ist wie allgemein üblich die informierte Einwilligung der Studienteilnehmer:innen erforderlich. Ist die Einholung einer solchen Einwilligung nicht umsetzbar, wie z. B. bei der Verknüpfung von Sekundärdaten, so muss die Einwilligung der Dateneigner und der Aufsichtsbehörden eingeholt werden. Will man Primärdaten mit Sekundärdaten verknüpfen wie etwa in der NAKO-Gesundheitsstudie (German National Cohort (GNC) Con-

sortium 2014), ist sowohl die Einwilligung der Teilnehmenden als auch der Dateneigner und Aufsichtsbehörden einzuholen (Stallmann et al. 2015). Dabei ist für jeden Einzelfall ein eigenes genehmigungspflichtiges Datenschutzkonzept vorzulegen. Da zudem an den verschiedenen Stellen in der Regel nicht dieselben Pseudonyme verwendet werden, ist die Einrichtung von Vertrauensstellen erforderlich, die eine entsprechende (De-) Pseudonymisierung vornehmen, bevor die Daten verknüpft und zur Auswertung weitergegeben werden können. Die Qualität der Verknüpfung hängt dabei unter anderem von den dafür eingesetzten personenidentifizierenden Variablen wie Name oder Krankenversichertennummer, der Güte dieser Variablen und dem verwendeten Record Linkage-Verfahren ab. Alternativ zur Verknüpfung der Daten über personenidentifizierende Variablen können verschiedene personenbezogene Datensätze auch über bestimmte individuelle Charakteristika verknüpft werden, was aber zu einer erhöhten Anzahl an falschen Verknüpfungen und so zu möglicherweise erheblichen Einschränkungen bezüglich der Qualität der Ergebnisse führen kann.

Gerade die Erforschung der Epidemologie von COVID-19 in Deutschland hat deutlich gemacht, dass die oben ausgeführten Möglichkeiten zum Record Linkage personenbezogener Gesundheitsdaten zu zeitaufwändig sind, um aus umfassenden Datenanalysen schnell informierte Entscheidungen ableiten zu können (s. auch die Stellungnahme der Interdisziplinären DFG-Kommission für Pandemieforschung 2021). Eine wichtige Voraussetzung, den Prozess zu beschleunigen, wäre die Einführung eines „unique identifiers", wie er z. B. in Dänemark als „central person registration (CPR)"-Nummer verwendet wird. Die CPR-Nummer ist in der Gesellschaft akzeptiert und in das Gesundheitssystem integriert, so dass alle Register für wissenschaftliche Zwecke über diese Nummer miteinander verknüpft werden können. Dabei verfügt Dänemark über mehr als 100 hochwertige Register und blickt auf eine lange Tradition in der Forschung mit Registerdaten zurück. Es überrascht daher nicht, dass Dänemark auch als „Data Heaven" bezeichnet wird (Holm, Ploug 2017). Die Datenhaltung erfolgt zentral in einem geschützten Bereich bei Statistics Denmark. Ein Antrag auf projektspezifischen Zugang muss an die dänische Datenschutzbehörde gestellt werden. Die anschließende Datenanalyse erfolgt ebenso in einem geschützten Bereich auf einem entsprechend der Forschungsfrage vorselektierten, verknüpften Datensatz (s. dazu auch das Beispiel in Abschnitt 2). Für die Verknüpfung mit Primärdaten ist wie in Deutschland eine informierte Einwilligung der Teilnehmenden erforderlich.

Ausblick

Derzeit erschweren starke rechtliche Restriktionen in Deutschland die Nachnutzung von personenbezogenen Daten zu Forschungszwecken. Im Sinne einer effizienten Ressourcennutzung zum Aufbau einer Nationalen Forschungsdateninfrastruktur müssen Lösungen gefunden werden, wodurch sich die datenschutzrechtlichen Rahmenbedingungen so gestalten lassen, dass sie einerseits moderne Forschung zulassen und andererseits die Interessen der/s Einzelnen schützen. In diesem Zusammenhang ist zu prüfen, ob gegebenenfalls auch durch den Einsatz von Methoden der Künstlichen Intelligenz die derzeitige Praxis eines Gastaufenthalts an einer datenhaltenden Institution, die Bereitstellung von Analysedatensätzen durch die datenhaltende Institution oder die kontrollierte Datenfernverarbeitung (Remote Access) vereinfacht und derart umgestaltet werden könnten, dass eine sichere Verknüpfung verschiedener personenbezogener Datensätze möglich ist. Zudem sollten für die jeweiligen Datenhalter Anreizsysteme für das Teilen bzw. Zugänglichmachen von Gesundheitsdaten geschaffen werden, denn außerhalb der rechtlichen Restriktionen sind auch die organisatorischen Hürden nicht zu unterschätzen.

Zusammenfassend lässt sich festhalten, dass noch viele Anstrengungen erforderlich sind, um das nachhaltige Teilen von Daten für die Forschung zu ermöglichen, aber auch, dass es dringend erforderich ist, alles dafür zu tun, denn: „There is a strong argument to be made that leaving data unshared is an impediment to the scientists of the future." (Nature Communications Editorial, 19. Juli 2018).

Danksagung

Ein Teil dieser Arbeit ist im Rahmen des NFDI4Health-Konsortiums entstanden. Wir danken der Deutschen Forschungsgemeinschaft (DFG) für die finanzielle Unterstützung – Projektnummer 442326535. Teile der in dieser Publikation beschriebenen Radiomics-Plattform werden im Rahmen der NFDI4Health Task Force COVID-19 entwickelt, mit Förderung durch die Deutsche Forschungsgemeinschaft (DFG, Projektnummer 45126528).

Literatur

Boulesteix AL, Wright MN, Hoffmann S, König IR. 2020. Statistical learning approaches in the genetic epidemiology of complex diseases. *Hum Genet* 139: 73–84.

Deist TM et al. 2017. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 4:24–31.

de Jong J et al. 2019. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience* 8, giz134.

de Jong J et al. 2021. Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. *Brain* 144:1738–1750

Fluck J et al. 2021. NFDI4Health – Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten. *Bausteine Forschungsdatenmanagement* 2:72–85.

Foraita R et al. 2018. Aufdeckung von Arzneimittelrisiken nach der Zulassung: Methodenentwicklung zur Nutzung von Routinedaten der gesetzlichen Krankenversicherungen. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 61:1075–1081.

Holm S, Ploug T. 2017. Big Data and health research – The governance challenges in a mixed data economy. *J Bioeth Inq* 14: 515–525.

Gemeinsame Wissenschaftskonferenz. 2018. Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018. https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf (Letzter Zugriff: 28.12.2021)

German National Cohort (GNC) Consortium. 2014. The German National Cohort: aims, study design and organization. *Eur J Epidemiol* 29: 371–382.

Gootjes-Dreesbach L, Sood M, Sahay A, Hofmann-Apitius M, Fröhlich H. 2020. Variational Autoenco der Modular Bayesian Networks for simulation of heterogeneous clinical study data. *Front Big Data* 3: 16.

Interdisziplinäre Kommission für Pandemieforschung der Deutschen Forschungsgemeinschaft (DFG). 2021. Daten für die gesundheitsbezogene Forschung müssen besser zugänglich und leichter verknüpfbar sein. https://www.dfg.de/download/pdf/foerderung/corona_infos/stellungnahme_daten_gesundheitsforschung.pdf (Letzter Zugriff: 30.12.2021).

Khanna S et al. 2018. Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci Rep* 8: 11173, doi: 10.1038/s41598-018-29433-3

Lessmann N et al. 2021. Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence. *Radiology* 298:E18–E28.

Linden T et al. 2021. An explainable multimodal neural network architecture for predicting epilepsy comorbidities based on administrative claims data. *Front Artif Intell* 4: https://www.frontiersin.org/articles/10.3389/frai.2021.610197/full

Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2021. https://www.nfdi.de/ (Letzter Zugriff: 28.12.2021).

Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten (NFDI4Health). 2021. www.nfdi4health.de (Letzter Zugriff: 28.12.2021).

Nature Communications Editorial. 2018. Data sharing and the future of science. *Nat Commun* 9: 28.

Overhoff D et al. 2021. The International Radiomics Platform – An initiative of the German and Austrian Radiological Societies – First application examples. *Rofo* 193: 276–288.

Rat für Informationsinfrastrukturen. 2016. Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Göttingen; https://rfii.de/download/rfii-empfehlungen-2016/ (Letzter Zugriff: 28.12.2021).

Schmidt CO et al. 2021. Die NFDI4Health – Task Force COVID-19. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 64: 1084–1092.

Sood M et al. 2020. Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders. *Sci Rep* 10: 10971.

Stallmann C et al. 2015. Individuelle Datenverknüpfung von Primärdaten mit Sekundär- und Registerdaten in Kohortenstudien: Potenziale und Verfahrensvorschläge. *Gesundheitswesen* 77: e37 – e42.

Taleb NN. 2021. The big errors of big data. https://fs.blog/the-big-errors-of-big-data (Letzter Zugriff: 28.12.2021).

Van Soest J et al. 2018. Using the Personal Health Train for automated and privacy-preserving analytics on vertically partitioned data. *Stud Health Technol Inform* 247: 581 – 585.

Watson DS, Wright MN. 2021. Testing conditional independence in supervised learning algorithms. *Mach Learn* 110: 2107 – 2129.

Wendland P et al. 2021. Generation of realistic synthetic data using multimodal neural ordinary differential equations. medRxiv, doi: https://doi.org/10.1101/2021.09.26.21263968

Wilkinson MD et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018

Wright MN, Kusumastuti S, Mortensen LH, Westendorp RGJ, Gerds TA. 2021. Personalised need of care in an ageing society: The making of a prediction tool based on register data. *J R Stat Soc Series A* 184:1199 – 1219.