Denkanstöße
aus der Akademie

11
Jan/2023

Eine Schriftenreihe der Berlin-Brandenburgischen
Akademie der Wissenschaften

Olaf Dössel, Tobias Schäffter, Britta Rutert (Hrsg.)

KÜNSTLICHE INTELLIGENZ IN DER MEDIZIN

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

KÜNSTLICHE INTELLIGENZ IN DER MEDIZIN

KÜNSTLICHE INTELLIGENZ IN DER MEDIZIN

Olaf Dössel, Tobias Schäffter, Britta Rutert (Hrsg.)

Denkanstöße 11 / Jan 2023

Informationen zur Publikationsreihe

In der Reihe „Denkanstöße aus der Akademie' werden Beiträge von Mitgliedern der Berlin-Brandenburgischen Akademie der Wissenschaften zu aktuellen forschungspolitischen und wissenschaftlichen Themen veröffentlicht. Die namentlich gekennzeichneten Beiträge geben die Auffassung der Verfasserinnen und Verfasser wieder. Sie repräsentieren nicht notwendigerweise den Standpunkt der Akademie als Institution.

Herausgeber: Der Präsident der Berlin-Brandenburgischen Akademie der Wissenschaften

Redaktion: Ute Tintemann
Grafik: Satz: eckedesign GmbH Berlin; Entwurf: angenehme Gestaltung/Thorsten Probst
Druck: PIEREG Druckcenter Berlin GmbH

Jägerstr. 22-23, 10117 Berlin, www.bbaw.de

ISBN: 978-3-949455-18-6

INHALTSVERZEICHNIS

HINTERGRUND UND ZUSAMMENFASSUNG

Olaf Dössel, Tobias Schäffter, Britta Rutert

Der Denkanstoß „KI in der Medizin" ist das Resultat eines gleichnamigen Symposiums, das am 4. Dezember 2021 an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) stattfand. Es ist der zweite „Denkanstoß", den die thematische Arbeitsgruppe „Digitalisierung der Medizin" der interdisziplinären Arbeitsgruppe (IAG) „Zukunft der Medizin: Gesundheit für alle" publiziert. Während sich die IAG allgemein mit zukünftigen Entwicklungen in der Medizin beschäftigte, befasste sich diese thematische Arbeitsgruppe mit der Frage, wie Kommunikationstechniken, künstliche Intelligenz, Big Data und Digitalisierung die Zukunft der Medizin beeinflussen und verändern werden.

Der erste „Denkanstoß" mit dem Titel „Apps und Wearables für die Gesundheit" (2021)[1] befasst sich dezidiert mit der Nutzung und dem Nutzen digitaler Gesundheitshelfer wie Handy-Apps und kleinen, ständig am Körper getragenen Geräten. Demgegenüber geht der vorliegende „Denkanstoß" auf den Einsatz von künstlicher Intelligenz (KI) und maschinellem Lernen in der Medizin ein. KI hält ein großes Potential für eine bessere medizinische Praxis und Forschung bereit. Die US-amerikanische Zulassungsbehörde (FDA) hat von 2016 bis 2020 bereits 29 KI-basierte Medizinprodukte zugelassen.[2] Die meisten dieser Technologien wurden für die Bereiche Radiologie, Kardiologie bzw. Allgemeinmedizin entwickelt. In den nächsten Jahren wird sich diese Zahl erheblich steigern, da viele neue Zulassungen angestoßen wurden. Gleichzeitig gilt es, zahlreiche technische, rechtliche und ethische Herausforderungen zu klären. In Europa zielt die neue EU-Verordnung zur KI („EU AI-Act") darauf ab, das breite Spektrum von KI-Anwendungen zu regulieren, um sie durch einen risikobasierten Ansatz mit den Werten und Grundrechten der EU in Einklang zu bringen. Auch eine Enquetekommission des deutschen Bundestages hat sich ausführlich mit dem Thema „Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale" beschäftigt.[3] Im Rahmen der „Plattform Lernende Systeme" wurde u. a. der Band „Sichere KI-Systeme für die Medizin – Datenmanagement und IT-

1 Dössel O, Schäffter T, Kutyniok G, Rutert B (Hrsg.). 2021. Apps und Wearables für die Gesundheit, Berlin (Denkanstöße aus der Akademie 7), https://edoc.bbaw.de/frontdoor/index/index/docId/3696

2 Benjamens S, Dhunnoo P, Meskó B. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med, 11(3): 118.

3 https://www.bundestag.de/webarchiv/Ausschuesse/ausschuesse19/weitere_gremien/enquete_ki

Sicherheit in der Krebsbehandlung" (2020) publiziert.[4] Auch vom Themennetzwerk Gesundheitstechnologien der Deutschen Akademie für Technikwissenschaften – acatech ist ein Positionspapier mit dem Titel „Machine Learning in der Medizintechnik – Analyse und Handlungsempfehlungen" (2020) erschienen.[5]

Insbesondere Fragen nach der Verantwortung, der Güte von Daten und Algorithmen und dem Schutz der individuellen Gesundheitsdaten sind in der Wissenschaft und der Öffentlichkeit fortwährend ambivalent diskutierte Themen. Das Symposium „KI in der Medizin" setzte sich auch mit dem Dilemma zwischen Datennutzen und Datenschutz in der Medizin auseinander und suchte nach möglichen Auswegen. Die zentrale Frage lautet dabei, wie digitale Daten für die wissenschaftliche Forschung sinnvoll und gewinnbringend genutzt werden können, ohne dabei den Datenschutz außen vor zu lassen oder ihm sogar einen so großen Raum zu geben, dass er wichtige Forschung verhindert. Auch werden ethische Fragen zum „richtigen" Umgang mit den Ergebnissen von KI in der Medizin diskutiert.

Der erste Teil dieses „Denkanstoßes" führt in das Thema KI in der Medizin und in die Anwendungsfelder ein. In dem vorliegenden Band schreiben dazu Olaf Dössel zu „Maschinelles Lernen und Künstliche Intelligenz in der Medizin – eine Einführung und ein Plädoyer", Fabian Schöck zu „Produktentwicklung und Zulassung von KI-Lösungen in der Medizinischen Bildgebung" sowie Christoph Lippert einen Beitrag über „Das Potential von Künstlicher Intelligenz für die Früherkennung von Krankheiten". Im zweiten Teil steht dann die Güte von Algorithmen und Software, von Daten und die Möglichkeiten zum Zugang zu vertrauenswürdigen medizinischen Daten im Vordergrund, und zwar mit Beiträgen von Tobias Schäffter (gemeinsam mit Daniel Schwabe und Stefan Haufe) zu „Gute Daten für eine vertrauenswürdige KI in der Medizin" und von Iris Pigeot (gemeinsam mit Holger Fröhlich, Timm Intemann, Guido Pause und Marvin N. Wright) zu „KI und die Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten (NFDI4HEALTH)". Auch der Artikel von Rico Barth (gemeinsam mit Peter Ganten und Manuela Urban) „Nachhaltige Medizin braucht digitale Souveränität" gehört zu diesem Themenbereich. Im dritten und letzten Teil kommen dann gesellschaftliche und ethische Fragen zur Sprache: Daniel Strech diskutiert in seinem Beitrag eine „Verantwortungsvolle Sekundärnutzung von Patientendaten" und Frauke Kreuter (gemeinsam mit Christoph Kern und Patrick Oliver Schenk) betrachtet in ihrem Artikel „Automatisierte Entscheidungen: Aspekte von Fairness, Datenqualität und Privacy".

4 https://www.acatech.de/publikation/sichere-ki-systeme-fuer-die-medizin/

5 https://www.acatech.de/publikation/machine-learning-in-der-medizintechnik/

Erkenntnisse

Jeder der acht Beiträge endet mit einem eigenständigen Resümee. Im Rahmen des Symposiums wurde nicht versucht, gemeinsame Ergebnisse abzuleiten. Daher folgen an dieser Stelle Erkenntnisse, welche die Organisatoren des Symposiums aus den Beiträgen gewonnen haben. Wenn im Folgenden Hinweise auf einzelne Artikel angegeben sind, so gibt das Gesagte nicht unbedingt die Meinung der jeweiligen Autor:innen wieder, sondern es soll die Leser:innen anregen, in den einzelnen Artikeln mehr dazu zu lesen.

Es gibt schon viele Anwendungen der KI in der Medizin. Viele davon sind erfolgreich durch die Prüfung nach der Medizingeräte-Verordnung (MDR) hindurchgegangen und wurden von den Prüfeinrichtungen als sicher und nützlich eingestuft (siehe Beiträge von OD, FS, TS, RB, IP). Auch bei der Früherkennung von Krankheiten und bei der Abschätzung von Risiko-Scores kann KI eine wichtige Rolle spielen (CL, IP).

Die meisten Anwendungen von KI in der Medizin ersetzen den Arzt nicht, sondern haben nur eine „beratende Funktion". Sie treffen selber keine Entscheidungen. Die Verantwortung für Entscheidungen verbleibt beim Arzt/der Ärztin (OD, CL). Hier besteht die Gefahr, dass Ärzt:innen den Empfehlungen des KI-Systems folgen, ohne sie kritisch zu bewerten und mit allen anderen Gesichtspunkten abzuwägen. Die Benutzerschnittstelle (user interface) eines KI-Systems müsste so konzipiert sein, dass der Benutzer/die Benutzerin so weit wie möglich gezwungen wird, noch einmal selber nachzudenken und die Vorschläge der KI-Anwendung kritisch zu prüfen.

Manchmal müssen KI-Systeme in der Medizin auch ohne einen Arzt/eine Ärztin Entscheidungen treffen, z. B. wenn in Notfallsituationen kein Arzt/Ärztin erreichbar ist (OD). Auch Screeningprogramme in Ländern ohne flächendeckende medizinische Versorgung erfordern manchmal Entscheidungen, die ohne einen Arzt/eine Ärztin auf der Grundlage begrenzter Informationen getroffen werden müssen (CL). Sie zu verbieten, kann dazu führen, dass Menschen sterben, deren Leben mit Hilfe des KI-Systems gerettet werden könnte.

In manchen Fällen sind die automatisierten Vorschläge des KI-Systems vom Benutzer/von der Benutzerin nicht sofort nachvollziehbar. Daher sollten KI-Systeme in der Medizin so weit wie möglich transparent sein und ihre Aussagen erklären können. Es gibt schon mehrere Ansätze, um diese Erklärbarkeit von KI-Systemen

zu erreichen. Sie machen aus der Black Box zunehmend eine Gray Box. Um hier zu besseren Methoden zu kommen, ist zunächst eine genauere Definition der Begriffe notwendig. Was heißt „erklärbar"? Welche Varianten von „Erklärbarkeit" müssen unterschieden werden? Kann die Transparenz bzw. Erklärbarkeit eines KI-Systems quantitativ bestimmt werden, um dem Nutzer ein Maß an die Hand zu geben, wie gut die Erklärbarkeit in einem KI-System erreicht wurde (OD, TS)? Hier besteht Forschungsbedarf.

Die heute und vielleicht zukünftig eingesetzten KI-Systeme in der Medizin sind sehr vielfältig und unterscheiden sich stark bezüglich der Risiken, die durch deren Einsatz auftreten können. Dies ist ähnlich zu bestehenden Risikoklassen von Medizinprodukten, deren Zulassung höhere Auflagen für höhere Risikoklassen fordert. Dazu sollten in der Diskussion über KI in der Medizin die Anwendungsgebiete präziser unterschieden werden. Es sollten nicht Urteile für ein Anwendungsbeispiel gefällt werden, die dann automatisch für alle KI-Systeme gelten sollen (OD).

Im Gegensatz zu klassischen Algorithmen, bei denen der Rechenweg definiert wird, werden in der KI andere Verfahren eingesetzt, die nicht notwendigerweise zu vorhersehbaren Ergebnissen führen. So werden oft Methoden verwendet, welche Zusammenhänge und Gesetzmäßigkeiten aus Daten lernen, um diese dann auf neue Daten anzuwenden. Mit falschen oder unvollständigen Daten gespeist, können solche Verfahren der KI aber auch falsche oder verzerrte Ergebnisse liefern. Für solche Fehler bei der Erzeugung des KI-Systems ist der Hersteller verantwortlich – er muss sie nach allen Regeln der Kunst ausschließen (Qualitätskontrolle). Dabei spielt die Güte und vor allem die Repräsentativität der Daten eine große Rolle. Letztere beschreibt, in wie weit genügend Daten für unterschiedliche Fälle verwendet wurden. Es gibt einen großen Forschungsbedarf die Fehlerrate einer KI-Methode aufgrund der Unsicherheit in den Daten zu charakterisieren, um so eine bessere Risikoabschätzung der KI-Entscheidung zu ermöglichen (TS).

Insgesamt sind Fehler aufgrund begrenzter Informationen über Patient:innen unvermeidbar. Dies gilt aber auch für die bestehende klinische Praxis. Daher sollte das KI-System den Nachweis erbringen, dass es weniger Fehler als ein Arzt/eine Ärztin macht (OD). Zum Nachweis und zur Bewertung von fehlerhaften Entscheidungen werden statistische Verfahren verwendet. Man untersucht dazu falsch positive und falsch negative Ergebnisse. Dabei kommt man automatisch zu der Frage, wann eine klinische Studie mit einer begrenzten Anzahl an Studienteilnehmern generalisierbar ist. Dies ist aber bei allen klinischen Studien problematisch,

insbesondere natürlich bei Studien mit Komponenten der KI: Sind die Ergebnisse auf alle Menschen, welche die Einschlusskriterien erfüllen, übertragbar? (CL, FK). Oft sind die guten Ergebnisse, die in einem ersten Test erreicht werden, in einem zweiten unabhängigen Test nicht reproduzierbar. In der evidenzbasierten Medizin werden bei der Festlegung neuer Leitlinien „Evidenzlevels" eingeführt. Eine einzige klinische Studie hat noch nicht genügend Aussagekraft, um eine neue Leitlinie zu begründen. Erst mehrere unabhängige Studien mit ähnlichem Ergebnis führen zu einem hohen Evidenzlevel.

Fast alle publizierten Beispiele von medizinischen Systemen mit KI basieren auf retrospektiven Studien. Manchmal werden prospektive Studien gefordert. Diese werden zum Nachweis des Erfolges einer Behandlung eingesetzt und sind dort die optimale Methode. Fast alle heute bekannten KI-Systeme in der Medizin sind aber rein diagnostisch: Die Entscheidung über die richtige Behandlung liegt weiter beim Arzt/der Ärztin. Daher kann man – ähnlich wie bei den bildgebenden Verfahren der Medizin – argumentieren, dass retrospektive Studien für den Nachweis der Wirksamkeit gut genug sind. Man erkennt aber, dass die Frage, wie der klinische Nachweis der Güte eines KI-Systems in der Medizin aussehen muss, noch nicht klar genug beantwortet werden kann. Hier besteht dringender Forschungsbedarf.

KI-Systeme in der Medizin müssen ein besonderes Augenmerk darauf legen, dass nicht bestimmte Patientengruppen bevorzugt und andere benachteiligt werden (OD, FK). Diese Forderung ist aus klassischen medizinischen Studien gut bekannt. Ethikkommissionen prüfen heute immer, ob eine beantragte medizinische Studie das Gebot der Fairness und der Nicht-Diskriminierung erfüllt. Dazu müssen klassische statistische Verfahren mit sozialwissenschaftlichen Betrachtungen kombiniert werden, um zu untersuchen, ob bestimmte Bevölkerungsgruppen in den Daten unter- oder überrepräsentiert sind (FK, TS). Insgesamt kann ein Bias in der Medizin nicht gänzlich ausgeschlossen werden – das gilt für die klassische Medizin und für die KI-unterstützte Medizin gleichermaßen. Allerdings wäre eine stärkere Transparenz der Datenlage wünschenswert, ähnlich wie es bei klinischen Studien gefordert und durchgeführt wird.

Zur Rechtsprechung bei Streitfällen im Bereich Arzthaftung – Patientenhaftung – Produkthaftung im Zusammenhang mit KI-Systemen müssen Leitlinien erarbeitet werden. Rechtsunsicherheit führt dazu, dass Unternehmen keine KI-Systeme anbieten werden, da sie hohe Schadensersatz-Forderungen befürchten müssen. Damit würde aber auch der Nutzen für viele Patienten blockiert.

Die Regeln des Datenschutzes und der Schutz der Privatsphäre müssen bei der Erstellung eines KI-Systems und beim Gebrauch eingehalten werden. Hierzu gibt es Gesetze, die selbstverständlich zu beachten sind (OD). Auch auf die Publikationen der Datenethikkommission zu diesem Thema soll hier hingewiesen werden.[6] Es gibt verschiedene Ansätze, wie Daten aus verschiedenen klinischen Studien zusammengeführt werden können, ohne die Datenhoheit des Patienten über seine Daten zu beeinträchtigen (CL, IG).

Mit dem Begriff „Sekundärnutzung" von Patientendaten ist die Nutzung der Daten für Forschungsprojekte gemeint, für die sie ursprünglich nicht erhoben und damit auch nicht explizit von den Patientinnen und Patienten freigegeben wurden. Viele Patient:innen – würde man sie fragen – hätten damit keine Probleme (FK). Datenschützer:innen verbieten eine solche sekundäre Datennutzung und bestehen darauf, dass die Patient:innen explizit um ihre Erlaubnis gebeten werden müssen. Kann eine unabhängige Prüfung der „Vertrauenswürdigkeit, Nützlichkeit und Einhaltung ethischer Rahmenbedingungen" des geplanten Forschungsvorhabens hier einen erleichterten Zugang zu Daten ermöglichen? Wie müsste solch eine Prüfung aussehen (DS)? Kann eine andere Perspektive auf die Problematik wie die „contextual integrity" das Problem auflösen (FK)?

Für die Prüfung von KI-Systemen müssen Regeln erarbeitet und als verbindliche Normen verabredet werden (OD, FS, TS). Die vom Bundesministerium für Wirtschaft und Klimaschutz initiierte „KI-Normungsroadmap" von DIN und DKE liefert wichtige Beiträge zu diesem Thema.[7] Insbesondere die Methoden zur objektiven Messung der Treffsicherheit und der Unsicherheit müssen besser definiert, vereinheitlicht und überall gleichermaßen angewendet werden. Das betrifft insbesondere die Messung der Güte der Algorithmen, welche stark mit der Güte der Datenbasis zusammenhängt, mit denen das KI-System trainiert wurde (TS). Qualitätstests mit Referenzdaten und mit sogenannten Benchmark-Tests können hier in Zukunft eine wichtige Rolle spielen (TS). Es sollte ein europaweit geltendes Qualitätszertifikat für KI in der Medizin geschaffen werden.

KI-Systeme in der Medizin sind nur auf der Grundlage großer und qualitativ hochwertiger Datensätze möglich (OD, FS, TS, IP). Die Initiativen zum Aufbau dieser

6 https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=E0CE0FB6C393186E4FBFB27917D9BB39.2_cid340?__blob=publicationFile&v=7

7 https://www.dke.de/resource/blob/2008010/0c29125fa99ac4c897e2809c8ab343ff/nr-ki-deutsch---download-data.pdf

Datenbanken müssen konsequent vorangetrieben werden (NFDI4Health siehe IP, GAIA-X for Health, Forschungsdatenzentrum am Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM)). Es muss besser geklärt werden, unter welchen Bedingungen Unternehmen Zugriff auf diese Daten bekommen. Schließt man die Industrie von der Nutzung grundsätzlich aus, so können viele Ergebnisse am Ende nicht für Produkte genutzt werden, welche Millionen von Patient:innen zu Gute kommen. Die „Datenspende" muss als etwas Gutes angesehen werden. Dazu werden in verschiedenen europäischen Ländern unterschiedliche Ansätze zur Freigabe der Daten für die Forschung durch den Patienten verfolgt („informierte Zustimmung"). In einigen Ländern müssen Patienten aktiv erklären, dass sie nicht mit einer Weiterverwendung einverstanden sind (opt-out), während in anderen eine explizite Zustimmung (opt-in) zur Nutzung eingefordert wird. Insgesamt sollte die informierte Zustimmung nicht so kompliziert sein, dass „normale" Patient:innen sie nicht mehr verstehen (OD, TS, FK).

Simulierte oder „augmentierte" Daten sind ein interessanter Weg, um zu einer großen Zahl von qualitativ hochwertigen Daten zu kommen, ohne dass die Rechte der Patient:innen betroffen sind (IP, TS, CL). Dies ist insbesondere in Hinblick auf Referenzdatensätze für eine unabhängige Prüfung der Treffsicherheit von KI-Systemen wichtig. Allerdings hat die Aussagekraft auch Einschränkungen: Geben die künstlich erzeugten Datensätze exakt die Wirklichkeit wieder?

„Open Data", „Open Source Software" und „Open Platforms" sind wichtige Methoden der wissenschaftlichen Forschung. Die „FAIR"-Prinzipien sollten so weit wie möglich auch bei Daten und Algorithmen der KI für die Medizin angewendet werden (RB, IP, DS). Das gilt insbesondere für Daten, die mit Hilfe von öffentlichen Mitteln gewonnen wurden (z. B. DFG- oder EU-geförderte Forschungsprojekte). Das weltweite Teilen der Gesundheitsdaten von Patienten stößt aber manchmal an die Grenzen des Datenschutzes. Die Anonymisierung der Daten ist nicht immer und vollständig möglich. Es ist potentiell denkbar, dass in Zukunft neue Technologien eine Re-Identifizierung von anonymisierten Daten erlauben könnten. Auch das Teilen von KI-Algorithmen stößt an Grenzen: Unternehmen müssen wirtschaftlich arbeiten. Es ist daher nicht sinnvoll, grundsätzlich zu fordern, dass Datensätze und Software, die mit großen wirtschaftlichen Vorleistungen eines Unternehmens erstellt wurden, publiziert werden müssen.

Eine wichtige ethische Problematik in der Medizin ist die Verteilung einer Ressource, die nur in begrenzter Zahl zur Verfügung steht (Verteilungsgerechtigkeit). Das könnte ein Beatmungsgerät für Corona-Patienten sein oder ein Spenderorgan für

Patienten, deren Leben nur mit einer Transplantation gerettet werden kann (FK). KI-Systeme könnten hier eingesetzt werden, um zu objektiven Empfehlungen zu kommen. Entscheidet dann ein KI-System über Leben und Tod? Sollte das verboten werden? Wichtig ist in diesem Zusammenhang die Beobachtung, dass solche schwierigen Entscheidungen in der Medizin auch schon vor der Erfindung der KI täglich gefällt werden müssen. Das Bundesverfassungsgericht hat vor einigen Monaten den Gesetzgeber aufgefordert, zum Schutz vor einer Benachteiligung wegen einer Behinderung bei einer Triage einen gesetzlichen Rahmen vorzugeben.[8] Selbstverständlich müsste ein KI-System so konzipiert sein, dass es diesem gesetzlichen Rahmen streng folgt.

Zusammenfassend erkennen die Organisatoren des Symposiums und Herausgeber dieses „Denkanstoßes" viele Anwendungen der KI in der Medizin, mit denen die Gesundheit von Menschen länger erhalten oder eine Krankheit besser geheilt werden kann. Es zeigt sich, dass dazu eine interdisziplinäre Zusammenarbeit von Experten aus Technik, Medizin, Ethik sowie Rechts- und Sozialwissenschaften notwendig ist um eine „vertrauenswürdige KI in der Medizin"[9] zu etablieren. Es werden einige offene Fragenfelder aufgeworfen, die durch Forschungsinitiativen geklärt werden müssen:

- Definition von Kriterien einer vertrauenswürdigen KI (Treffsicherheit, Transparenz, Erklärbarkeit, Fairness etc.);

- Bereitstellung von Vergleichstests für KI-Methoden;

- Regeln für notwendige klinische Studien zum Nachweis des klinischen Nutzens von KI-Methoden und Klärung der Frage, wann die Ergebnisse einer Studie generalisierbar sind;

- Prozeduren zur Förderung von Vertrauenswürdigkeit, Nützlichkeit und Ethik von (KI-) Studien;

- Festlegung von Kriterien, wann die letzte Entscheidung bei einem Arzt/einer Ärztin verbleiben muss und wann ein KI-System auch eigenständig handeln darf;

8 BVerfG, Beschluss des Ersten Senats vom 16. Dezember 2021 – 1 BvR 1541/20 –, Rn. 1–131, http://www.bverfg.de/e/rs20211216_1bvr154120.html

9 https://www.medizin.nrw/news/pruefkatalog-vertrauenswuerdige-kuenstliche-intelligenz/

- Rechtssicherheit für Patient:innen, Ärzt:innen und Unternehmen im Falle von Streitfällen;

- Anpassung regulatorischer Rahmenbedingungen für eine Zulassung und Zertifizierung von KI-Medizinprodukten;

- Bereitstellung einer Dateninfrastruktur zur breiteren Nutzung von primären und sekundären Forschungsergebnissen als Grundlage einer vertrauenswürdigen KI;

- Entwicklung von Standards der notwendigen Datenqualität für vertrauenswürdige KI-Methoden;

- Definition von Kriterien zur informellen Selbstbestimmung für sichere und faire Datennutzung, welche die Forschung mit Gesundheitsdaten nicht unmöglich machen;

- Entwicklung und Bereitstellung „offener Referenzdaten" und „offener KI-Referenzalgorithmen";

- Erhöhung der gesellschaftlichen Akzeptanz von KI-Methoden in der Medizin durch Partizipation von Gesunden, Patient:innen und Ärzt:innen.

MASCHINELLES LERNEN UND KÜNSTLICHE INTELLIGENZ IN DER MEDIZIN – EINE EINFÜHRUNG UND EIN PLÄDOYER

Olaf Dössel

Sowohl Patient:innen als auch Ärzt:innen haben sehr unterschiedliche Auffassungen zu den Themen maschinelles Lernen (ML) und künstliche Intelligenz (KI) in der Medizin. Dieser Artikel soll ein Beitrag zur Versachlichung sein. Er beginnt mit einer Einführung in das Thema. Nach einer Abgrenzung werden vier Beispiele vorgestellt, wo ML und KI schon heute in der Medizin eingesetzt werden. Das Thema Apps und Wearables im Zusammenhang mit ML und KI wird kurz beleuchtet. Dann werden Aspekte zu den folgenden Themen beschrieben: mögliche Fehler, Verantwortung, ethische Fragen, Zulassung, Normen und Gesetze. Auch wird die Frage gestellt, wie mehr qualitativ hochwertige und annotierte Daten für die Forschung gewonnen werden können. Der Artikel endet mit Thesen und einem Plädoyer für einen klugen Umgang mit maschinellem Lernen in der Medizin.

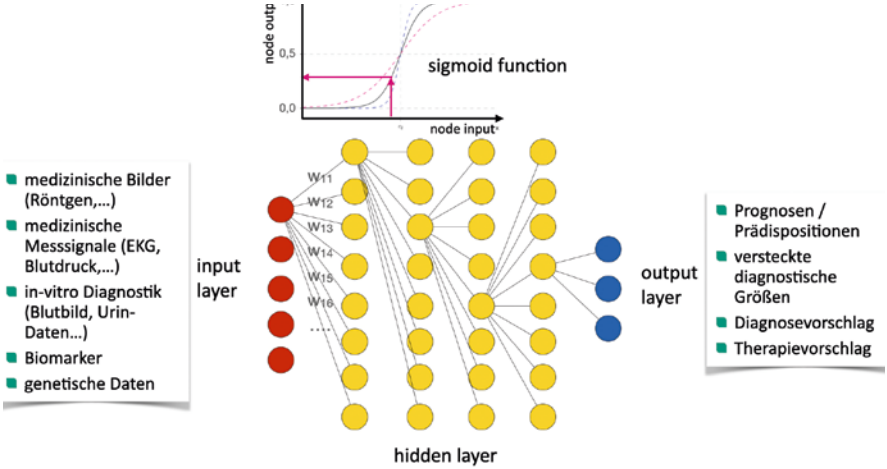


Abb. 1: Grundschema von maschinellem Lernen in der Medizin

Der „input layer" ist eine Folge von Zahlen, z. B. das Elektrokardiogramm. Der „output layer" ist ebenfalls eine Reihe von Zahlen, z. B. steht 1 für „Vorhofflimmern" und 0 für „kein Vorhofflimmern". Die gelben Kreise sind sogenannte „Knoten" (nodes). Der Eingang eines Knotens ist eine gewichtete Summe der Zahlen in der Schicht davor, die w_{ij} sind dabei die Gewichte. Der Ausgang des Knotens wird über eine Sigmoid-Funktion berechnet, bei der eine Steilheit und ein Schwellwert festgelegt werden. Beim Trainieren werden die Gewichtsfaktoren so bestimmt, dass möglichst viele Eingangsdaten auf den richtigen Ausgang abgebildet werden.

Abbildung 1 zeigt plakativ, wovon dieser Artikel handelt: Wir füttern einen Algorithmus mit medizinischen Daten wie z. B. Bildern, physiologischen Messsignalen, In-Vitro-Messwerten und genetischen Daten, aus denen Prognosen, Prädispositionen, versteckte diagnostische Größen, Diagnose- und Therapievorschläge generiert werden können. Der Artikel konzentriert sich also auf maschinelles Lernen und insbesondere auf das „supervised learning", bei dem ein Trainingsdatensatz mit gut bekannten Paaren aus Eingangs- und Ausgangsdaten zur Verfügung steht. Abbildung 1 visualisiert ein Deep Neural Network. Aber es gibt noch mehr Methoden des maschinellen Lernens: Polynomial-Regression, Support Vector Machines, Decision Trees und Random Forests, k-nearest neighbor und andere mehr.

Die meisten Publikationen zu diesem Thema sind heute im Bereich Bildverarbeitung angesiedelt. Ein wichtiges Beispiel ist die Befundung von Mammographie-Aufnahmen. 2017 nahmen in Deutschland 2,8 Millionen Frauen am Mammographie-Screening teil. Vorgeschrieben ist heute noch ein „Double Reading", d. h. zwei Expert:innen müssen sich jedes Bild ansehen und entscheiden, ob es Hinweise auf einen bösartigen Brusttumor gibt. Die Frage ist: Kann man dieses Verfahren durch „Single Reading" plus „Computer Assisted Diagnosis (CAD)" ersetzen oder vielleicht sogar verbessern? Dazu gibt es eine ganz neue Publikation vom September 2021 mit dem klaren Ergebnis: Single Reading plus CAD ist auf jeden Fall nicht schlechter als Double Reading.[1]

Ein ganz anderes Beispiel ist die Apple-Watch. Dabei geht es um die Frage, ob bei einem Anwender der Armbanduhr ein Verdacht auf Vorhofflimmern festgestellt werden kann. Dazu findet man eine Publikation im New England Journal of Medicine mit dem Titel „Large Scale Assessment of a Smart Watch to Identify Atrial Fibrillation".[2] 419.000 Menschen haben an der Studie teilgenommen. Die Apple-Watch konnte sehr erfolgreich Vorhofflimmern bei Patient:innen erkennen. Damit konnte gezeigt werden, was maschinelles Lernen in der medizinischen Diagnose ermöglichen kann.

1 Graewingholt A & Duffy S. 2021. Retrospective comparison between single reading plus an artificial intelligence algorithm and two-view digital tomosynthesis with double reading in breast screening, J Med Screen 28(3): 365–368. doi: 10.1177/0969141320984198

2 Perez MV, Kenneth W, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, Hung G, Lee J. 2019. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation, N Engl J Med; 381: 1909–1917, DOI: 10.1056/NEJMoa1901183

Das dritte Beispiel ist eine App, die verspricht, dass sie zuverlässig bösartige Melanome von gutartigen braunen Flecken auf der Haut unterscheiden kann.[3] Diese App hat dank maschinellen Lernens das Potential, Menschenleben zu retten. Ein zu spät erkannter Hautkrebs kann oft nicht mehr geheilt werden. Gleichzeitig kann Leben gefährdet werden, wenn der Benutzer sich auf einen falschen negativen Befund verlässt.

Ein letztes Beispiel: In Deutschland versterben jedes Jahr über 100.000 Menschen an Kammerflimmern. Der automatische Defibrillator kann das Leben dieser Menschen retten. Allerdings muss das Gerät zuverlässig erkennen, ob es sich um ein schockpflichtiges EKG handelt – schnell, zuverlässig und ohne einen Arzt[4]. In der nächsten Generation wird das noch besser gelingen – mit maschinellem Lernen.[5]

Man erkennt das extrem weite Spektrum an Anwendungen von ML in der Medizin. Die folgende Übersicht ist sicherlich nicht vollständig:

- **Diagnostische Bildgebung:** Trennen von Bildern mit und ohne Befund, Markieren auffälliger Bereiche, Registrieren, Segmentieren
- **Onkologie:** Identifikation der spezifischen Tumorklasse, genetische Eingruppierung
- **Strahlentherapie:** Vorschlag eines Bestrahlungsplans
- **Anästhesie:** Vorhersage kritischer Kreislaufzustände
- **Intensiv-Medizin:** Erkennung lebensbedrohlicher Zustände, Alarme
- **Ophthalmologie:** Befundung von OCT-Bildern

3 Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM & Thrun S. 2017. Dermatologist-level classification of skin cancer with deep neural networks, Nature, volume 542: 115–118.

4 Zur besseren Lesbarkeit werden in diesem Artikel abwechselnd die männliche und weibliche Form verwendet.

5 Shen C, Khedraki R, Cohoon T, Barakat A, Elashery AR, Shah K, Freed B, Walters D, Gibson DN, Olson N, Ng FS, Perry P, Rogers J and Bhavnani S. 2020. Development of a convolutional neural network for shockable arrhythmia classification within a next generation automated external defibrillator, Journal of the American College of Cardiology. Volume 75, Issue 11 Supplement 1, DOI: 10.1016/S0735-1097(20)34095-X

- K**ardiologie:** Auswertung von Langzeit-EKGs, z. B. Identifikation von Extrasystolen, Erkennen von Phasen mit Vorhofflimmern

- **Endokrinologie:** Ratgeber für die Insulingabe bei Patienten mit Diabetes mellitus

- **Dermatologie:** Erkennung maligner Melanome

- **Neurologie:** Vorhersage epileptischer Anfälle, Anfallstagebuch

- **Pathologie:** Erkennung suspekter Gebiete oder Befundung histologischer Schnitte von Gewebe

- **Notfall-Medizin:** Triage (schnelle Einschätzung, welche Patientin sofort ärztliche Hilfe benötigt und welcher Patient noch etwas warten kann), automatischer Defibrillator

In einem Positionspapier der acatech[6] wird folgendes Ordnungsschema vorgeschlagen:

- **Gruppe 1:** ML-Applikationen für den Laien ohne Einbeziehung eines Arztes. Beispiele: Smart Watch, Wearables, Apps für Laien.[7]

- **Gruppe 2:** Einteilung von Patienten in Gruppen unterschiedlicher Priorität. Beispiel: Mammographie Screening.

- **Gruppe 3:** Unterstützung der Ärztin bei der Diagnose und Therapieauswahl. Beispiel: Anästhesie-Assistent, Computer Assisted Radiology, hier entscheidet am Ende immer noch eine Ärztin.

- **Gruppe 4:** medizinische Anwendungen, ohne dass ein Arzt dabei ist. Beispiel: automatischer Defibrillator.

6 Acatech (Hrsg.). 2020. Machine Learning in der Medizintechnik. Analyse und Handlungsempfehlungen: 10–13, https://www.acatech.de/publikation/machine-learning-in-der-medizintechnik/

7 Siehe auch Dössel O, Schäffter T, Kutyniok G, Rutert B (Hrsg.). 2021. Apps und Wearables für die Gesundheit: Chancen und Herausforderungen", Denkanstoß 7, Berlin, https://edoc.bbaw.de/frontdoor/index/index/docId/3696

Quer dazu verläuft die Einteilung in zwei Fälle: (a) das Lernen ist beim Ausliefern des Medizinproduktes abgeschlossen (möglicherweise erfolgt nach einer gewissen Zeit ein Update) oder (b) das System passt sich auch nach der Auslieferung kontinuierlich an den Patienten an (selbstlernendes System). So könnte sich ein Assistent für den Diabetiker bei der Beratung für die Insulinabgabe kontinuierlich an den Patienten anpassen. Dem Autor sind keine zugelassenen Medizinsysteme vom Typ (b) bekannt. Trotzdem spielen sie in der öffentlichen Diskussion eine dominante Rolle.

Es folgt ein ganz kurzer Blick auf die Apps, die Smart Watches und die Wearables, also Anwendungen der Gruppe 1. Die gute Nachricht ist, dass erfolgreiche Apps dem Hersteller Millionen von Daten liefern, mit denen man hervorragend ML-Algorithmen trainieren kann. Damit sind auch gute Ergebnisse zu erwarten. Die schlechte Nachricht ist demgegenüber, dass die Erfassung der Daten oft unzuverlässig ist und dass in den Daten viele Fehler auftreten können. Zudem erfährt der Hersteller sehr persönliche Gesundheitsdaten vom Patienten, die er wiederum geschäftlich nutzen kann.

Im folgenden Abschnitt geht es um Fehler, die ein ML-Algorithmus machen kann. Es gibt zwei ganz unterschiedliche Arten von Fehlern, und es ist wichtig, dass diese Fehler in der Diskussion immer deutlich auseinandergehalten werden.

Es gibt Fehler bei der Implementation: Der Trainingsdatensatz enthält nicht ausgewogen alle Menschen, auf die der Algorithmus später angewendet werden soll (Bias), er enthält zu viele falsche Klassifizierungen, die Zahl der Trainingsdaten ist zu klein, der Algorithmus ist zu sehr an den Trainingsdatensatz angepasst und nicht generalisierbar (Overfitting und Leakage).

Auf der anderen Seite gibt es Fehler aufgrund einer unsicheren Datenlage: die „Wahrheit“ [8] ist zum Zeitpunkt, zu dem eine Entscheidung getroffen werden muss, nicht genau bekannt. Viele Entscheidungen in der Medizin basieren auf einer unsicheren Datenlage. Medizinische Studien zeigen immer eine statistische Schwankung. Es gibt fast keine Aussagen in der Medizin, die sich im Nachhinein immer als richtig erweisen.

8 Wahrheit ist hier gemeint im Sinne von Ground Truth, also im Sinne einer „bestmöglichen Entscheidung“.

Zum Fehler der ersten Art ist zu sagen: Die Qualität, die der Hersteller verspricht, muss er auch im klinischen Alltag halten können. Sonst haftet der Hersteller. Zum Fehler der zweiten Art ist zu sagen: Solche Fehler werden passieren. Das ist heute nicht anders. Diese Fehler treten bei allen Entscheidungen in der Medizin – mit oder ohne ML – auf. Aber wenn ein ML-Algorithmus nachweisen kann, dass er statistisch bessere Entscheidungen fällt als jede einzelne Ärztin, ist die Zahl der Fehler – gemittelt über alle Patienten – mit dem ML-Algorithmus kleiner als ohne ihn.

Kommen wir zum Thema Verantwortung als einem im Zusammenhang mit maschinellem Lernen besonders häufig genannten Begriff. Wer trägt bei der Verwendung von Systemen der Medizintechnik welche Verantwortung? Wir haben bereits ein sehr differenziertes System von Verantwortung und Haftung im Bereich der Medizin. Es ist fast immer eine Dreiecksbeziehung zwischen dem Patienten, der Ärztin, dem Hersteller. Jeder trägt für seinen Teil die Verantwortung. Die Patientin muss zu ihrer Behandlung eine „informierte Zustimmung" geben. Darüber hinaus gelten die Arzthaftung und die Produkthaftung. Der Hersteller muss so weit wie möglich sicherstellen, dass sein Produkt die Normen erfüllt und „hält, was es verspricht". Der Hersteller und sein Produkt werden von einer sogenannten benannten Stelle überwacht. Die benannten Stellen wiederum müssen sich gegenseitig überwachen. Täglich kommt es zu Streitfällen zwischen Patient:innen und Ärzt:innen. Der Patient meint: Meine falsche Behandlung hätte eine gute Ärztin besser gemacht. Die Ärztin entgegnet: Die „richtige" Behandlung konnte man zu dem Zeitpunkt gar nicht kennen. Meistens endet das mit dem Satz: Der unglückliche Ausgang war „schicksalshaft", ein in diesem Zusammenhang häufig verwendeter Begriff. Dann übernimmt tatsächlich niemand die Verantwortung für den schlechten Ausgang.

Es stellt sich nun die Frage, was genau bei Medizinprodukten, die ML-Algorithmen enthalten, eigentlich anders ist? Die meisten der oben genannten Aspekte gelten unverändert. Ein besonders wichtiger Aspekt ist aus meiner Sicht, dass die benannten Stellen prüfen, ob bei einem Medizinprodukt alle Normen eingehalten werden. Hier sehe ich Herausforderungen. Wir müssen möglicherweise die Normen anpassen, so dass die benannten Stellen klare Richtlinien bekommen, was sie bei einem ML-Medizinsystem genau prüfen müssen. Hierfür benötigen wir:

- Begriffsbestimmungen und genaue Definitionen,
- Anforderungen an die Qualität und die Größe der Datenbasis,
- Anforderungen an die Trennung von Trainings-, Validierungs- und Testdaten,
- Anforderungen an den Ausschluss/die Vermeidung von Bias,

- Anforderungen an die Angaben zur Treffsicherheit, zu Ausreißern und zur Robustheit,
- Anforderungen an die Transparenz,
- Anforderungen an die Interaktion mit der Ärztin (user interface).

Auf das Thema „Vermeidung von Bias", also der unbewussten Bevorzugung bzw. Benachteiligung von Bevölkerungsgruppen, wird später noch genauer eingegangen.

Der Begriff „Treffsicherheit" erfordert eine genauere Definition. Bei einer Klassifikationsaufgabe können generell die Sensitivität und die Spezifität[9] bestimmt werden. Weiter kann man im Einzelfall fragen, mit welcher Wahrscheinlichkeit eine andere Klassifizierung herausgekommen wäre (Klasse A: 51 % und Klasse B: 49 % oder Klasse A: 10 % und Klasse B: 90 %). Oder man kann fragen, ob eine leicht veränderte Eingangsgröße (z. B. durch einen Messfehler) zu dem gleichen oder zu einem ganz anderen Ergebnis geführt hätte. Diese letzten beiden Angaben sind für einen Arzt wichtige Informationen. Sie sollten immer angegeben werden, auch weil sie dem Arzt deutlich vor Augen führen, dass keine mit ML erzeugte Klassifikation oder Vorhersage mit 100 % Sicherheit richtig ist (Anforderung an das User Interface).

Auch der Begriff „Transparenz" muss genauer betrachtet werden. Man könnte Angaben darüber fordern, welches Merkmal in den Eingangsdaten zu der Entscheidung geführt hat. Einige der ML-Algorithmen (z. B. Entscheidungsbäume (Decision Trees)) können schon heute diese Information liefern. Bei den neuronalen Netzen ist das schwieriger. Daher spricht man dort oft von einer Black Box. Mathematiker haben aber Methoden erarbeitet, mit denen auch bei neuronalen Netzen aus der Black Box eine Gray Box wird, d. h. die Entscheidung des Algorithmus wird teilweise nachvollziehbar gemacht.[10] Diese Transparenz ist natürlich sehr wünschenswert. Sie sollte, wo immer möglich, dargestellt werden. Sie darf aber keine zwingende Voraussetzung sein, da diese Forderung nicht immer erfüllt werden kann. Übrigens basieren auch heute viele Leitlinien der Medizin nicht auf genau bekannten Ursache-Wirkung-Prinzipien, sondern auf statistischem Wissen aus klinischen Studien (Grundprinzip der evidenzbasierten Medizin).

9 https://de.wikipedia.org/wiki/Beurteilung_eines_binären_Klassifikators#Sensitivität_und_Falsch-negativ-Rate

10 Binder A, Montavon G, Bach S, Müller KR, Samek W. 2016. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layer, arXiv:1604.00825 or https://doi.org/10.48550/arXiv.1604.00825

Die an ML angepassten Normen müssen EU-weit gelten, besser noch weltweit und sie müssen nachprüfbar sein. Dabei stellt sich immer die Frage nach Referenzdatensätzen. Das sind Datensätze, die unabhängig vom Hersteller generiert werden und mit denen so objektiv wie möglich die Treffsicherheit quantitativ bestimmt werden kann. Der Vorschlag der EU zu einer „Verordnung zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz"[11] plädiert für eine risikobasierte Festlegung von Mindeststandards. Auch in Deutschland wird an einer „Normungsroadmap künstliche Intelligenz"[12] gearbeitet, die auch den Bereich „KI in der Medizin" umfasst.

Noch ein kurzer Blick auf die Gesetzeslage: Es sind in der 19. Legislaturperiode des Bundestages sehr viele Gesetze verabschiedet worden, die das Thema KI in der Medizin berühren. Insbesondere das Patientendatenschutzgesetz[13] und die digitale Gesundheitsanwendungen Verordnung[14]. Im Patientendatenschutzgesetz ist auch die elektronische Patientenakte (ePA) geregelt, und das hat Bezug zu unserem Thema, denn mit Hilfe der ePA können wichtige Daten für die Forschung generiert werden. Das Gesetz sieht eine freiwillige Datenspende für die Forschung ab 2023 vor. Hierauf soll später noch einmal genauer eingegangen werden.

Im nächsten Abschnitt soll kurz auf ethische Aspekte eingegangen werden. Es gibt mittlerweile sehr viele Publikationen zum Thema Ethik und KI in der Medizin.[15] Viele wiederholen nur Aspekte, die schon einmal vorher von anderen beschrieben wurden. Viele beschäftigen sich auch mit Problemen, die wir eigentlich gar nicht haben.

11 https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF, 2021.

12 https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf, 2020.

13 Patientendatenschutzgesetz PDSG 2020. https://www.bundesgesundheitsministerium.de/patientendaten-schutz-gesetz.html

14 Digitale Gesundheitsanwendungen Verordnung 2020. https://www.bundesgesundheitsministerium.de/service/gesetze-und-verordnungen/guv-19-lp/digav.html

15 Siehe z. B. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, in Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E. 2018. Minds and Machines. 28: 689–707, https://doi.org/10.1007/s11023-018-9482-5

Worauf wir bei ML in der Medizin achten müssen, ist völlig unbestritten [16]:

- Vorrang menschlichen Handelns und menschliche Aufsicht
- Technische Robustheit und Sicherheit
- Schutz der Privatsphäre und Datenqualitätsmanagement
- Transparenz
- Vielfalt
- Nichtdiskriminierung und Fairness,
- Gesellschaftliches und ökologisches Wohlergehen
- Rechenschaftspflicht

Die Themen „Technische Robustheit", „Transparenz" und „Rechenschaftspflicht" (Verantwortung) wurden schon oben ausführlich betrachtet. Der „Schutz der Privatsphäre" ist in der Europäischen Datenschutz-Grundverordnung und in der dazu gehörenden deutschen Datenschutz-Grundverordnung (DSGVO) [17] sehr stringent geregelt. Es ist selbstverständlich, dass auch Medizinprodukte, die Algorithmen des maschinellen Lernens enthalten, diese Gesetze einhalten müssen. Auch bei der Entwicklung solcher Produkte müssen diese Gesetze beachtet werden.

Zum Thema „Vielfalt und Nicht-Diskriminierung": Jede klinische Studie muss heute „Einschlusskriterien" und „Ausschlusskriterien" festlegen. Eine klinische Studie durchzuführen, die alle Menschen einschließt, ist fast unmöglich. Die Ergebnisse der Studie dürfen nur für die Patienten angewendet werden, die die Einschlusskriterien erfüllen. Dabei passieren auch Fehler – das war (leider) schon immer so. Das bedeutet für Medizinprodukte, die ML beinhalten: Eine ML-Anwendung muss genau spezifizieren, wofür und für wen sie gut ist (Medical Device Regulation (MDR) [18]). Im Datensatz und bei der klinischen Studie müssen alle Patient:innen, die die Einschlusskriterien erfüllen, gleichmäßig vertreten sein. Auf Gruppen von Patient:innen, die nicht beteiligt waren, darf das ML-Programm nicht angewendet werden. Das bedeutet zwar nicht, dass das Problem mit der Nichtdiskriminierung und Fairness (Bias) abschließend gelöst ist, aber es bedeutet, dass das Problem in der Medizin schon seit langem bekannt ist und es schon viele Richtlinien dazu gibt.

16 High-Level Expert Group on Artificial Intelligence of the European Commission (AI-HLEG), Ethics Guidelines for Trustworthy AI, 08.04.2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

17 Datenschutz-Grundverordnung (DSGVO), 2018, https://dsgvo-gesetz.de

18 https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32017R0745&from=DE

Zum Schluss noch ein Blick auf das Thema „Daten für die Forschung". Zurzeit werden drei Säulen der Datengewinnung für die medizinische Forschung aufgebaut: NFDI4Health[19], Gaia-X[20] und das Forschungsdatenzentrum am BfArM, das hier ausschließlich betrachtet werden soll.

Das Patientendatenschutzgesetz[21] regelt die Möglichkeit, Daten aus der elektronischen Patientenakte (ePA) für die Forschung freizugeben. Die gematik[22] hat den Auftrag, bis 2023 die Schnittstellen von der ePA zum Forschungsdatenzentrum am BfArM zur Verfügung zu stellen, das die Weitergabe der Daten für die Forschung organisiert. Die Daten aus der ePA können dann – nach der Einverständniserklärung der Patienten – schnell für Forschungsprojekte zur Verfügung stehen („Datenspende"). Das ist insbesondere aus dem Blickwinkel der Forschung und Entwicklung für neue medizinische Systeme mit maschinellem Lernen sehr zu begrüßen, da so hochwertige Datensätze mit Millionen von Patientendaten entstehen können. Aber wie kompliziert wird diese Einverständniserklärung („broad consent") sein? Und wer darf Anträge auf Datenzugriff beim BfArM stellen? Hier gibt es noch offene Fragen. Im Patientendatenschutzgesetz steht folgendes: „Das Forschungsdatenzentrum [...] macht die übermittelten Daten [...] folgenden Nutzungsberechtigten zugänglich: den Hochschulen, den nach landesrechtlichen Vorschriften anerkannten Hochschulkliniken, öffentlich geförderten außeruniversitären Forschungseinrichtungen und sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung, sofern die Daten wissenschaftlichen Vorhaben dienen". Eine offene Frage ist, ob die medizintechnische Industrie und die Pharma-Industrie im Verbund mit einer Universität Zugriff auf die Daten bekommen darf? Insgesamt haben wir in Deutschland einen Flickenteppich von Landesdatenschutzverordnungen und Krankenhausgesetzen, die in unterschiedlicher Weise mal das eine erlauben und das andere verbieten. Und die Patienteneinwilligung ist offenbar ein sehr kompliziertes Thema. Der sogenannte „broad consent", der von Expertinnen und Experten der Medizininformatikinitiative erarbeitet und zunächst von allen 16 Datenschutzbeauftragten für gut befunden wurde,[23] ist jetzt wieder umstritten.

19 https://www.nfdi4health.de

20 https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Artikel/UseCases/framework-of-medical-records-in-europe.html

21 Patientendatenschutzgesetz PDSG https://www.bundesgesundheitsministerium.de/patientendaten-schutz-gesetz.html

22 https://www.gematik.de/anwendungen/e-patientenakte/

23 https://www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung, 2020.

Eine Reihe von Umfragen zu diesem Thema zeigt, dass es in der Bevölkerung eine hohe Bereitschaft gibt, Gesundheitsdaten anonymisiert für die wissenschaftliche Forschung freizuschalten.[24] Aber meistens können die Daten nur pseudonymisiert werden, da es irgendwo einen Schlüssel gibt, mit dem die Identität des Datenspenders zurückgewonnen werden kann. Ohne diesen Schlüssel können Datenbanken nicht kontinuierlich um neue Daten der Patient:innen erweitert werden. Es gibt Bedenken in der Bevölkerung, Daten aus der klassischen Gesundheitsversorgung für kommerzielle Anbieter freizuschalten. Aber wie sollen die Ergebnisse der Forschung beim Patienten ankommen, wenn wir kommerzielle Anbieter ausschließen? Und bei Gesundheits-Apps geben viele Menschen ihre Daten völlig bedenkenlos kommerziellen Anbietern.

Zum Schluss sollen neun Thesen, die zum Teil aus der acatech-Arbeitsgruppe[25] hervorgegangen sind, vorgestellt werden.

- Maschinelles Lernen in der Medizin wird den Arzt nicht ersetzen, sondern ihn unterstützen.

- In wenigen Ausnahmen, zum Beispiel in Notfällen, ist ein sofortiges Handeln – auch ohne Anwesenheit einer Ärztin – durch das System mit maschinellem Lernen erforderlich.

- Medizinsysteme mit maschinellem Lernen sollten – wo immer möglich – erklären können, warum sie zu einer bestimmten Aussage gekommen sind („Transparenz"), und eine „Treffsicherheit" angeben.

- Die Prinzipien von Arzt- und Produkthaftung müssen in ihrer Anwendung auf maschinelles Lernen in der Medizin genauer untersucht werden. Rechtssicherheit ist eine notwendige Voraussetzung für die Einführung von ML in der Medizin.

- Die Forschungsförderung im Bereich des maschinellen Lernens in der Medizin wird als „gut" eingeschätzt. Trotzdem sind weitere Anstrengungen nötig, zum Beispiel in den Bereichen Transparenz, Treffsicherheit, Evaluierung und klinische Anwendung.

24 https://www.pwc.de/de/gesundheitswesen-und-pharma/healthcare-barometer-2020.html, 2020.

25 https://www.acatech.de/publikation/machine-learning-in-der-medizintechnik/, 2020.

- Der Schutz der persönlichen Daten muss – wie durch die Datenschutz-Grundverordnung (EU-DSGVO), das Bundesdatenschutzgesetz (BDSG) und das Patientendaten-Schutz-Gesetz (PDSG) vorgeschrieben – gewährleistet sein.

- Es sind Lösungen notwendig, um große medizinische Datenbanken für die Forschung und Entwicklung zusammenzustellen. Die Generierung von Referenzdatensätzen für die objektive Qualitätsmessung von ML-Algorithmen sollte gefördert werden. Die Einrichtung verschiedener Datenzentren für die Forschung wird begrüßt. Die Datenspende muss als „etwas Gutes" und für alle Menschen Nützliches angesehen werden.

- Es muss geklärt werden, wie klinische Studien für den Beweis der Wirksamkeit eines ML-Systems aussehen müssen.

- Die regulatorischen Aspekte der Zulassung und Zertifizierung von Medizinprodukten (Normen) müssen an neue Aspekte von ML angepasst werden.

Alle diese Überlegungen führen zu einem Plädoyer für einen klugen und besonnenen Einsatz von maschinellem Lernen in der Medizin. Die großen Vorteile sollten genutzt werden, ohne dabei die berechtigten Bedenken zu vernachlässigen.

PRODUKTENTWICKLUNG UND ZULASSUNG VON KI-LÖSUNGEN IN DER MEDIZINISCHEN BILDGEBUNG

Fabian Schöck

Zusammenfassung

In der medizinischen Bildgebung gibt es eine stetig wachsende Lücke zwischen der Nachfrage nach Befundung der Bilder und der Verfügbarkeit von Fachpersonal. Gleichzeitig gibt es bahnbrechende Entwicklungen im Bereich der KI-basierten Befundungsunterstützung, auch aufbauend auf der Verfügbarkeit großer Datenmengen, welche für das Training der KI-Algorithmen verwendet werden können. Daraus ergibt sich ein großes Potential für neue Software-Medizinprodukte, deren Einsatz Qualitätssprünge und eine effizientere Befundung erhoffen lässt. Gleichzeitig ergibt sich ein Paradigmenwechsel bei der Entwicklung solcher Produkte im Vergleich zu der klassischen Entwicklung bildgebender Modalitäten, da die Verfügbarkeit, Selektion und Güte von relevanten klinischen Daten zu den entscheidenden Erfolgsfaktoren werden. Neben dem Zugang zu solchen Daten bedeutet das streng regulierte Umfeld eine weitere, hohe Hürde für einen erfolgreichen Markteintritt. Es gibt dabei regional deutlich unterschiedliche Regelungen, die für erfolgreiche globale Lösungen bereits zu Beginn einer aufwändigen Entwicklung zu beachten sind. Siemens Healthineers hat eine lange Historie in der Forschung und Entwicklung von KI-Lösungen in der medizinischen Bildgebung. Mit der Produktfamilie AI-Rad Companion konnten für verschiedene Körperregionen, Modalitäten und Anwendungszwecke KI-Lösungen auf den Markt gebracht und ein beständig wachsendes Portfolio entwickelt werden, was einen Beitrag zu besseren klinischen Ergebnissen und zur Bewältigung der Herausforderungen in der medizinischen Bildgebung leisten soll.

Künstliche Intelligenz in der medizinischen Bildgebung – bereit für die klinische Routine?!

„AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't." (Curtis Langlotz, Stanford University)

Der Bedarf an medizinischer Bildgebung ist in den letzten Jahren beständig gestiegen und ein weiteres deutliches Wachstum ist zu erwarten. Gleichzeitig gibt es in den meisten Ländern einen erheblichen Nachwuchsmangel an Radiologen,

so dass sich eine rapide wachsende Lücke zwischen dem Bedarf an radiologischen Befundungen und deren Verfügbarkeit ergibt. Nach Prognosen des Royal College of Radiologists wird sich diese Lücke im Jahr 2023 schon auf mehr als 30 % der benötigten Arbeitszeit in der Radiologie belaufen.[1] Neben der reinen Anzahl der Untersuchungen spielt dabei auch eine wesentliche Rolle, dass die Bildgebung mit Volumendatensätzen, z. B. die Computertomographie (CT) und Magnetresonanztomographie (MRT), besonders stark zunimmt. Das hohe Aufkommen an zu befundenden Untersuchungen und die daraus resultierende hohe Arbeitsbelastung kann dramatische Konsequenzen für die Qualität der klinischen Arbeit haben. So zeigen Studien, dass retrospektiv ermittelte Fehlerraten bei der Befundung radiologischer Datensätze bei 30 % liegen.[2] Und bei der Untersuchung der Auswirkung von erhöhtem zeitlichen Druck auf die Fehlerrate hat sich gezeigt, dass eine Halbierung der Zeit für die Befundung zu einem Anstieg der Fehlerrate um 17 % geführt hat.[3] Diese Untersuchungen verdeutlichen schon heute den massiven Bedarf an Unterstützung für die Radiologen, und die Prognosen bezüglich der wachsenden Versorgungslücke vergrößern noch einmal die Hoffnung auf positive Verbesserungen, die durch den Einsatz von KI-Software erreicht werden können.

Im Hinblick darauf ist es verständlich, dass es regelmäßig ein großes Medienecho gibt, wenn KI-basierte Software in der medizinischen Bildgebung in wissenschaftlichen Untersuchungen gut abschneidet. Anhand verschiedener Beispiele aus der aktuellen Literatur soll dabei eine Einordnung sehr vielversprechender Ergebnisse, aber auch der kritische Blick auf die Umsetzung in der klinischen Realität gegeben werden.

Brustkrebs-Screening und Mammographie

In einer vielbeachteten Studie haben Rodriguez-Ruiz et al. (2019)[4] gezeigt, dass eine KI-Lösung in einem Vergleich mit 101 Radiologen besser als 61,4 % der Radio-

1 The Royal College of Radiologists. 2019. Clinical radiology, UK workforce census 2018 report. London: The Royal College of Radiologists.

2 Berlin L. 2007. Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol 189(3): 517–522. doi: 10.2214/AJR.07.2209. PMID: 17715094

3 Berlin L. 2015. Faster Reporting Speed and Interpretation Errors: Conjecture, Evidence, and Malpractice Implications. J Am Coll Radiol. 12(9): 894–896. doi: 10.1016/j.jacr.2015.06.010. PMID: 26355199.

4 Rodríguez-Ruiz A et al. 2019. Detection of breast cancer with mammography: effect of an artificial intelligence support system, Radiology 290.2: 305–314.

logen abschnitt, quantifiziert durch die AUC.[5] In Summe wurden im Rahmen der Studie 28.296 unabhängige Befundungen durchgeführt. Dabei konnte gezeigt werden, dass in diesem retrospektiven Rahmen die KI-Lösung einer Befundung durch einen Menschen mindestens gleichwertig war.

Tuberkulose-Screening und Thoraxröntgen

In einem Vergleich von fünf verschiedenen KI-Lösungen zur Detektion von Tuberkulose auf Röntgenthorax-Bildern konnte gezeigt werden, dass alle fünf Software-Lösungen den erfahrenen und zertifizierten Radiologen überlegen waren.[6] Gleichzeitig konnte mit dem Einsatz der KI-Lösungen eine hohe Sensitivität im Screening erhalten werden. Der Einsatz der KI sorgte auch dafür, dass aufwändige weitere molekulare *Xpert*-Tests ohne negative Auswirkung auf den Erfolg des Screening-Programms reduziert werden konnten. Unter anderem diese Studie führte dazu, dass die WHO im März 2021 die Empfehlung aussprach, im Tuberkulose-Screening KI-Lösungen als Standard zu etablieren.[7]

Bestimmung des Knochenalters in der Kinderradiologie

Eine prospektive Multi-Center-Studie von Eng et al. (2021) konnte nachweisen, dass durch den Einsatz einer KI-Lösung sowohl die Genauigkeit bei der Bestimmung des Knochenalters im Vergleich zur Befundung ohne KI-Lösung signifikant erhöht werden konnte und dass gleichzeitig auch die Zeit für die Befundung bedeutend kürzer war.[8] Gerade durch den prospektiven Ansatz der Studie, an der verschiedene unabhängige Krankenhäuser teilnahmen, wurde die Studie viel beachtet. So wurde unter anderem in einem Editorial in der Zeitschrift *Radiology*

5 AUC: Area Under the Receiver Operating Curve; das gängige Qualitätsmaß zur Bewertung eines Vergleichs verschiedener KI-Systeme oder von KI-Systemen gegenüber Befundung durch einen Menschen.

6 Qin ZZ et al. 2021. Tuberculosis detection from CXR for triaging in a high tuberculosis-burden setting: evaluation of five artificial intelligence algorithms. The Lancet Digital Health. doi: 10.1016/S2589-7500(21)00116-3

7 World Health Organization. 2021. WHO consolidated guidelines on tuberculosis. Module 2: screening systematic screening for tuberculosis disease. In WHO consolidated guidelines on tuberculosis. Module 2: screening systematic screening for tuberculosis disease: 6868.

8 Eng DK et al. 2021. Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment: A Prospective Multicenter Randomized Controlled Trial. Radiology. 2021 Dec; 301(3): 692–699. doi: 10.1148/radiol.2021204021. Epub 2021 Sep 28. PMID: 34581608.

von David A. Rubin von der NYU Grossman School of Medicine der wegweisende Charakter dieser Studie für das sich ideal ergänzende Zusammenspiel der menschlichen Befundung und den standardisierten, effizienten Ergebnissen einer qualitativ hochwertigen KI-Lösung hervorgehoben.[9]

Detektion von Lungenrundherden mit Thoraxröntgen

Eine wichtige Multi-Center-Studie im Bereich des Thoraxröntgens haben Radiologen der Harvard Medical School, der Ludwig-Maximilians-Universität München und dem MVZ Prof. Uhlenbrock & Partner, Dortmund, zusammen mit Siemens Healthineers veröffentlicht. Die Studie zeigte auf, dass der unterstützende Einsatz der KI-Lösung *AI-Rad Companion Chest X-ray* die Genauigkeit und AUC-Performance der Radiologen signifikant verbesserte.[10] Die Studie unterschied dabei auch zwischen Radiologen mit wenig und viel Erfahrung und konnte nachweisen, dass für beide Kategorien eine signifikante Verbesserung der Befundungsergebnisse erzielt werden konnte.

KI-Lösungen als Heilmittel für die klinische Routine?

Die vorgestellten Studien zeigen eindrucksvoll, dass in den letzten Jahren rasant KI-Lösungen entwickelt wurden und dass für einige klinische Fragestellungen die Qualität der menschlichen Befundung bereits erreicht oder sogar übertroffen werden kann. Die Anwendung von KI-Lösungen in der breiten klinischen Routine bedarf jedoch noch weiterer Untersuchungen, gerade auch im Hinblick auf eine Generalisierung und die spezifischen Einsatzszenarien. So untersuchten beispielsweise Freeman et al. (2021) in einem systematischen Review die Studienlage für den Einsatz von KI-Lösungen im Mammographie-Screening.[11] Die Studie weist darauf hin, dass die Evidenz derzeit noch nicht hinreichend ist, um die im Mammo-

9 Rubin DA. 2021. Assessing Bone Age: A Paradigm for the Next Generation of Artificial Intelligence in Radiology. Radiology 301(3): 700–701.

10 Homayounieh F, Digumarthy S, Ebrahimian S, Rueckel J, Hoppe BF, Sabel BO, Conjeti S, Ridder K, Sistermanns M, Wang L, and Preuhs A. 2021. An Artificial Intelligence–Based Chest X-ray Model on Human Nodule Detection Accuracy from a Multicenter Study. JAMA network open, 4(12): e2141096-e2141096

11 Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A and Taylor-Phillips S. 2021. Use of artificial intelligence for image analysis in breast cancer screening programs: systematic review of test accuracy. Bmj: 374.

graphie-Screening übliche Doppelblind-Befundung durch eine einfache, aber KI-unterstützte Befundung zu ersetzen. Die Autorinnen und Autoren fordern die wissenschaftliche Community und die Hersteller der KI-Lösungen gleichermaßen dazu auf, die Anzahl großer, prospektiver, randomisierter Multi-Center-Studien mit großen Patient:innenkollektiven deutlich auszuweiten, um die Validierung für den Einsatz in der klinischen Routine zu verbessern.

Ein Paradigmenwechsel – die Entwicklung KI-basierter Medizinprodukte

Dieser kurze Einblick in die Studienlage zeigt, welche große Bedeutung die klinische Validierung der Qualität der KI-Lösungen für den Erfolg eines Medizinprodukts hat. Für die Produktentwicklung führt das zu einem Paradigmenwechsel in Bezug auf vier wesentliche Aspekte:

1. *Zweckbestimmung („intended purpose")*
   Bei KI-Lösungen in der Radiologie handelt es sich um neue Medizinprodukte ohne eine entsprechende Historie von Vorgängerprodukten. Die präzise Zweckbestimmung ist daher das Fundament zu Beginn der Produktentwicklung und bestimmt für den weiteren Verlauf zentrale Fragen rund um die Auswahl passender Daten, die Integration in die IT-Systeme der Nutzer und die Interaktion zwischen Menschen und KI-Software („Workflow Integration"). Insbesondere der letzte Aspekt besitzt eine zentrale Bedeutung, da die KI-Algorithmen in der klinischen Routine in den seltensten Fällen isoliert Ergebnisse liefern, sondern in komplexe radiologische/klinische Arbeitsabläufe integriert werden und der Mehrwert der Lösungen dort im Zusammenspiel zwischen KI und Mensch garantiert sein muss.

2. *Datenstrategie*
   Abgeleitet aus der Zweckbestimmung und den angestrebten quantitativen Qualitätskriterien (z. B. Sensitivität, Spezifität, AUC-Werte) muss zu Beginn eine Strategie im Hinblick auf die benötigten Daten und ihre Verfügbarmachung ausgearbeitet werden. Dabei spielen Aspekte der Datenannotation, der Zusammensetzung der Daten im Hinblick auf die (internationale und diverse) Patientenpopulation und viele weitere Faktoren eine Rolle. Im Sinne des geflügelten Wortes „garbage in, garbage out" ist es für eine gute KI-Lösung unverzichtbar, dass die Trainings- und Validierungsdaten zum einen den höchsten klinischen Standard garantieren, zum anderen aber auch mit den jeweiligen etablierten Befundungsstandards vereinbar sind. Nur so sind

einwandfreie Ergebnisse gesichert, welche auch die Akzeptanz in der klinischen Routine finden.

3. *Regulatorische Strategie*
   Ebenfalls zu Beginn des Entwicklungsprozesses muss klar definiert werden, in welchen Ländern und unter welchen Regularien das Produkt in den Verkehr gebracht werden soll. Dabei sind die teilweise erheblichen Unterschiede bei der Zulassung und die unterschiedliche Kategorisierung von Medizinprodukten in verschiedenen Regionen der Welt zu berücksichtigen, was im nächsten Abschnitt noch in weiteren Details erläutert werden soll.

4. *Klinische Validierung*
   Die Studien aus dem vorherigen Abschnitt haben gezeigt, dass die Zusammenarbeit mit einem breiten Spektrum an erfahrenen klinischen Partner:innen ein wesentlicher Faktor für die erfolgreiche Entwicklung von KI-Lösungen ist. Sowohl in der Frühphase der Produktdefinition als auch in der Validierung vor einer Freigabe des Medizinproduktes spielen diese Partner:innen eine zentrale Rolle. Die Strategie für die klinische Validierung muss daher verschiedene Faktoren vereinen. Das bedeutet unter anderem, dass verschiedene Nutzerprofile abgedeckt werden müssen, eine regionale Abdeckung für alle geplanten Länderzulassungen zu berücksichtigen ist, und gleichzeitig auch, dass das Patient:innenkollektiv repräsentiert ist, für das die Lösung entwickelt wird.

Für eine erfolgreiche KI-Produktentwicklung müssen diese vier Aspekte frühzeitig bedacht und zu einem Gesamtkonzept verbunden werden. Dieses Konzept beinhaltet die verschiedenen Phasen der Entwicklung, und zwar angefangen bei der frühen Innovation über die weitere Entwicklung von Prototypen und Produkt-Software bis hin zur Phase der Zulassung und des kommerziellen Rollouts der Software. In der Produktentwicklung bei Siemens Healthineers hat es sich als erfolgreich erwiesen, möglichst frühzeitig über Prototypen das Feedback der Nutzer:innen einzuholen, dann in kurzen Zyklen neue Funktionalitäten oder Verbesserungen auszurollen und kontinuierlich mit den Nutzer:innen zu validieren. Auf der Entwicklungsseite wird das mit einem agilen Entwicklungsmodell ermöglicht, das auf den Prinzipien von DevOps und kontinuierlicher Integration und Deployment beruht. Die frühe Phase der Datensammlung, -aufbereitung und -selektion ist, wie erwähnt, für den Erfolg eines KI-Medizinprodukts entscheidend. Die relevanten Bereiche werden bei Siemens Healthineers daher gebündelt, und zwar unter dem Begriff der „Data Factory". Ein zentrales Element ist dabei das „Big Data Office".

In einem klar definierten Rahmen und unter zentraler Governance werden hier die für die Entwicklungsprojekte benötigten Daten gesammelt, aufbereitet und verwaltet. Über klar definierte Prozesse erhalten die Data-Scientist-Teams für die jeweilige Entwicklung Zugriff auf die relevanten Daten und sind damit in der Lage, die passenden KI-Modelle zu trainieren und zu testen. Dafür gibt es eine Supercomputing-Infrastruktur, auf der mehr als 1.000 KI-Experimente pro Tag gefahren werden. Mit diesem Ansatz lässt sich eine beachtliche Skalierung erreichen, so dass im letzten Geschäftsjahr insgesamt fast 150 parallele Projekte bearbeitet werden konnten. Bei mittlerweile fast 70 Lösungen ist KI-Software aus dieser „Fabrik" ein wichtiger Bestandteil.

Die regulatorische Strategie als wichtiges Erfolgskriterium

Bevor KI-Lösungen, wie die oben beispielhaft genannten, in die klinische Routine gebracht werden können, gilt es noch die jeweiligen Länderzulassungen als Hürde zu nehmen. Die nationalen Regularien für Medizinprodukte sind ein komplexes Feld und die jeweiligen Behörden sind weit von einer einheitlichen Handhabung für KI-basierte Medizinprodukte entfernt. Der Aufwand und die Zulassungszeit für ein Medizinprodukt hängen sehr stark davon ab, wie die Klassifizierung im jeweiligen Land erfolgt und wie umfassend die klinischen Claims sind. Teilweise gibt es in den Regularien global sehr uneinheitliche oder sogar sich widersprechende Regelungen. So gibt es in den USA unter den FDA-Regularien eine Klasse von „Computer Aided Triage and Notification Software", welche einfacher und schneller zugelassen werden kann als klassische „Computer Aided Detection Software". Demgegenüber sieht die EU Medical Device Regulation (MDR) die Klasse IIb vor, wenn die Software unter die Bestimmung „Triage/Drives clinical management" fällt. Regelungen in weiteren Ländern wie z. B. China bedeuten für globale Hersteller eine noch größere Komplexität, auch im Hinblick auf die klinische Validierung der KI-Software, welche unter Umständen im jeweiligen Land erfolgen muss. Hinzu kommt, dass viele der relevanten Regelungen weltweit erst im Entstehen sind. Auch die nationalen Behörden sind noch dabei, die richtigen Ansätze zu finden, um schnelle Innovation zu fördern, ohne aber gleichzeitig die strengen Maßstäbe für eine Medizinprodukte-Zulassung aufzuweichen. Die US-Behörde FDA hat für die Weiterentwicklung der Zulassung von KI-Lösungen im Oktober 2021 die „Guiding Principles for Good AI/ML Practice" veröffentlicht und damit auch die SW-Hersteller zur Debatte und gemeinsamen Gestaltung sinn-

voller Regelungen eingeladen.[12] Und auch seitens der EU-Kommission gibt es entsprechende Initiativen und Veröffentlichungen, zum Beispiel für die Beurteilung vertrauenswürdiger KI-Lösungen.[13] Es lässt sich also zusammenfassen, dass für eine erfolgreiche KI-Lösung zum einen eine gute globale Strategie vonnöten ist, zum anderen aber das Verständnis und die lokale Kompetenz zu jeweiligen nationalen Regularien gefragt ist.

Die Klassifizierung und die verschiedenen Anwendungsbeispiele von KI wurden bereits genannt. An dieser Stelle soll noch kurz erwähnt werden, welche Arbeitsschritte von KI unterstützt bzw. übernommen werden können und wie sich das in eine höhere regulatorische Komplexität übersetzt. Die einfachsten KI-Lösungen unterstützen zum Beispiel dabei, die Reihenfolge in der Radiologie-Worklist zu ändern, so dass besonders relevante Fälle gesondert ausgewiesen werden und zeitnah befundet werden können. Das ist das, was im FDA-Sprachgebrauch als „Triage Software" oder CADt bezeichnet wird. Komplexere Lösungen können dann relevante Befunde genau lokalisieren, quantifizieren und auch eine genaue Charakteristik vorschlagen, z. B. ob es sich bei einem auffälligen Rundherd im Thorax-CT mit erhöhter Wahrscheinlichkeit um einen malignen Tumor handelt. Daraus lassen sich im Rahmen der klinischen Guidelines dann auch die Empfehlungen für die weitere Diagnostik oder sogar die Therapie ableiten. In der höchsten Kategorie würden solche Lösungen dann eigenständig ohne menschliche Befundung operieren. Regulatorisch zählen diese zu den Lösungen, die ein „rule out CAD" anbieten. Das heißt, es handelt sich hier um Lösungen, die ohne menschliche Befundung schon nicht relevante, normale Fälle aussortieren und damit signifikant zu einer schnelleren, effizienteren Radiologie beitragen könnten. Allerdings gibt es für solche Ansätze verständlicherweise auch entsprechend hohe medico-legale Hürden, so dass es bisher noch keine KI-Lösung in der Radiologie mit einer entsprechenden Zulassung gibt.

12 https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

13 https://2021.ai/eu-takes-on-trustworthy-ai-with-altai/

AI-Rad Companion – eine wachsende Produktfamilie von KI-Anwendungen in der Bildgebung

Für den abschließenden Teil soll kurz der AI-Rad Companion, also die Familie von KI-Lösungen, vorgestellt werden, mit denen Siemens Healthineers für verschiedene Anwendungen in der Radiologie und Radioonkologie eine führende Rolle auf dem globalen Markt einnimmt. Die verschiedenen AI-Rad-Companion-Lösungen decken unterschiedliche bildgebende Modalitäten und Körperbereiche ab. Ziel ist es, das medizinische Fachpersonal so zu unterstützen, dass die Befundung effizienter erfolgt. Gleichzeitig soll auch die Qualität gesteigert werden, zum Beispiel im Hinblick auf Zufallsbefunde oder schwierige klinische Fragestellungen. Die AI-Rad Companion-Softwarelösungen werden über die Teamplay Digital Health Platform in die IT-Landschaft der Nutzer:innen integriert, wobei die Ausführung der KI-Algorithmen über verschiedene Ansätze mit Cloud- oder Edge-Computing erfolgen kann. Abbildung 1 zeigt schematisch, wie die Bilddaten von der Modalität zur Berechnung an AI-Rad Companion verschickt werden. Die Berechnung kann entweder lokal oder in der Cloud erfolgen. Die verschiedenen Funktionalitäten der Chest CT-Lösung sind unterhalb des Bildes dargestellt, es werden zum Beispiel automatisierte Messungen vorgenommen, relevante Strukturen hervorgehoben, quantifiziert und in einem übersichtlichen Bericht zusammengefasst. Sobald die Ergebnisse vorliegen, werden sie zum Routine-Arbeitsplatz weitergeleitet und dort integriert in der Befundungssoftware dargestellt.

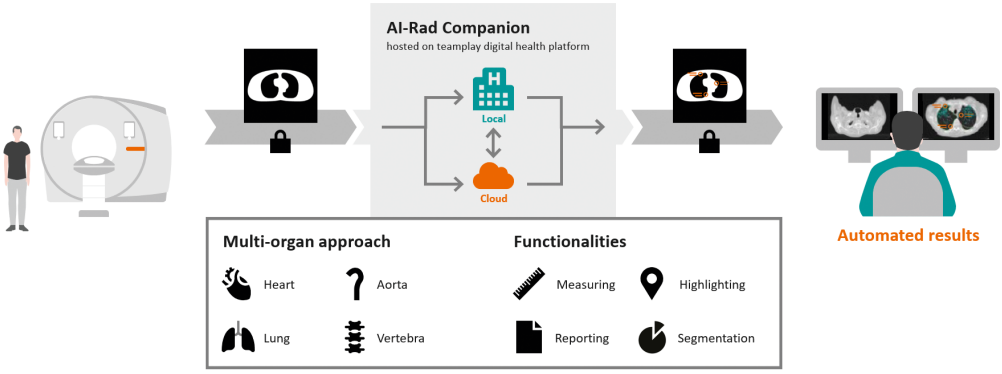


Abb. 1: Schaubild zum Fluss der Daten und zur Funktionalität des AI-Rad Companion Chest CT.

Aktuell erstreckt sich das AI-Rad Companion-Portfolio auf fünf verschiedene Lösungen:

- Brain MR (Unterstützung in der Kopf-MR-Bildgebung)
- Chest CT (Unterstützung für die Befundung von Thorax-CT, vor allem im Hinblick auf Lungenrundherde, kardiovaskuläre und muskuloskelettale Fragestellungen)
- Chest X-ray (Unterstützung für die Befundung von Thoraxröntgenbildern)
- Prostate MR (Unterstützung im wachsenden Segment der MR-Bildgebung für die Prostata)
- Organs RT (Unterstützung für automatisierte Konturierung/Segmentierung in der Strahlentherapie)

Weitere Lösungen, mit denen das Spektrum der Körperregionen und der klinischen Fragestellungen noch deutlich erweitert werden soll, sind derzeit in Arbeit. Darüber hinaus liegt ein großer Fokus auf der kontinuierlichen Weiterentwicklung der bestehenden Lösungen hin zu noch umfassenderer Unterstützung der klinischen Nutzer und besserer Performance der Algorithmen.

Neben der Verbreiterung des klinischen Spektrums der AI-Rad Companion-Produktfamilie konzentriert sich unsere Arbeit auch auf die nahtlose Einbindung in den klinischen Arbeitsablauf, d. h. auf die Workflow-Integration. Um diese zu garantieren, liefert AI-Rad Companion die Ergebnisse in den gängigen Austauschformaten, und Siemens Healthineers engagiert sich stark in internationalen Gremien wie IHE oder die DICOM Working Groups, welche die Definition und Umsetzung solcher Standards vorantreiben. Auch Aspekte der technischen Performance bei der Berechnung der Ergebnisse, die Sicherheit der Patientendaten und der Schutz vor Cyber-Angriffen spielen eine wichtige Rolle bei der Akzeptanz der KI-Lösungen auf dem Markt.

Eine beständig steigende Zahl klinischer Kunden weltweit setzt die AI-Rad Companion-Lösungen bereits ein und in vielen Studien konnte der klinische Mehrwert bestätigt werden. So zeigen Rueckel et al. (2021) für die CT-Befundung

in der Notaufnahme, dass durch den Einsatz der KI-Software die Anzahl von übersehenen Befunden deutlich reduziert werden kann.[14]

Ausblick

Die dargelegten Beispiele zeigen, dass KI-Lösungen in der medizinischen Bildgebung bereits beachtliche Erfolge im Vergleich mit der menschlichen Befundung erzielen können. Die Grundlage für den Erfolg der KI-Lösungen sind Faktoren, welche eine immense weitere Skalierung in den nächsten Jahren versprechen:

- Qualitativ hochwertige Daten als Grundlage für das Training der KI-Algorithmen werden in immer größerer Anzahl und für viele verschiedene klinische Fragestellungen verfügbar,

- Rechenleistung und Speicherplatz für Hardware-intensives Training der KI-Modelle werden beständig günstiger und erlauben somit eine sich sicher noch beschleunigende Weiterentwicklung neuer KI,

- deutliche Fortschritte von KI-Lösungen in der nicht-medizinischen Bildgebung im Endkunden-Markt der Unterhaltungselektronik sorgen für positive Effekte auch bei der Entwicklung von Lösungen speziell für die medizinische Bildgebung,

- die Digitalisierung in Krankenhäusern und Radiologien ermöglicht eine nahtlose Einbindung von KI-Ergebnissen in den Workflow der Befundung und damit eine höhere Effizienz bei gleichzeitigem Qualitätsgewinn.

Das Vertrauen in das große Potential zeigt sich auch in der Bewertung des Markts für KI-Firmen im Bereich der medizinischen Bildgebung. Allein im Bereich der Startups bewegen sich die Investments in KI-Firmen bereits im Milliardenbereich.[15] Viele Firmen und deren Investoren sind davon überzeugt, mit ihren KI-Lösungen

14 Rueckel J, Sperl JI, Kaestle S, Hoppe BF, Fink N, Rudolph J, Schwarze V, Geyer T, Strobl FF, Ricke J and Ingrisch M. 2021. Reduction of missed thoracic findings in emergency whole-body computed tomography using artificial intelligence assistance. Quantitative Imaging in Medicine and Surgery, 11(6): 24–86.

15 Alexander A, Jiang A, Ferreira C and Zurkiya D. 2020. An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. Journal of the American College of Radiology 17(1): 165–170.

in den nächsten Jahren und Jahrzehnten in den Routinemarkt vordringen zu können.

Die verschiedenen hier im Artikel skizzierten Punkte bei der Entwicklung und Freigabe der KI-Lösungen müssen beachtet werden, um nachhaltigen Erfolg und eine starke Position im Wettbewerb sicher zu stellen. Von zentraler Bedeutung ist und bleibt dabei der Faktor Mensch. Es muss das Ziel sein, die etablierten, hohen Qualitätsstandards in der Radiologie und das große Potential für klinischen Mehrwert der KI-Lösungen ideal zu kombinieren, und so eine bessere Patientenversorgung und eine Effizienzsteigerung gleichzeitig zu realisieren.

DAS POTENTIAL VON KÜNSTLICHER INTELLIGENZ FÜR DIE FRÜHERKENNUNG VON KRANKHEITEN

Christoph Lippert

Die Analyse von multiparametrischen Gesundheitsdaten hat das Potential, die Gesundheitsfürsorge zu revolutionieren, hin zu einem System, in dem wir vielen Krankheiten durch vorbeugende Maßnahmen und frühe Interventionsmaßnahmen gezielt in einem Frühstadium begegnen. Dies ist möglich, da pathologische Prozesse in der Regel lange vor dem Auftreten klinischer Symptome beginnen. Ziel der Früherkennung ist es, Krankheitsprozesse zu identifizieren, bevor sie sich in klinischen Symptomen manifestieren. Die Identifizierung von Individuen mit erhöhtem Krankheitsrisiko ist ein essenzieller Bestandteil der Früherkennung, da sie einen effizienteren Einsatz von präventiven Maßnahmen erlaubt. So besteht beispielsweise in Familien, in denen schon mehrere Frauen an Brustkrebs erkrankt sind, eine erhöhte genetische Prädisposition für Brustkrebs, so dass in diesen Familien ein entsprechendes Screening verstärkt eingesetzt werden sollte, um Tumore in einem Frühstadium zu entdecken.

Im Gegensatz zur traditionellen Gesundheitsversorgung, in der einer ärztlichen Behandlung meist ein Krankheitsverdacht und Krankheitssymptome vorausgehen, sollten Screeningverfahren auch bei asymptomatischen Individuen einsetzbar sein. Um eine regelmäßige Anwendung zuzulassen, sollten diese möglichst keine Belastung für den Patienten/die Patientin darstellen. Hier besitzen nichtinvasive Verfahren, u. a. basierend auf Bildgebung und Genetik, ein großes Potential. Auf Künstlicher Intelligenz (KI) basierende Verfahren werden sowohl entwickelt, um beispielsweise anhand genetischer Daten das persönliche Risiko für Krankheiten einzuschätzen als auch für den Einsatz im Screening, beispielsweise durch Computerassistierte Detektionsverfahren (CAD) für die automatisierte Bildanalyse. Biobanken und medizinische Repositorien speichern bereits einen großen Bestand an verschiedenen Arten von Genomik- und Gesundheitsdaten, die das Entwickeln von KI-Modellen für die Risikoschätzung und für die Früherkennung ermöglichen. Die integrative Analyse dieser Daten, einschließlich Genetik, Genomik, Bilddaten, tragbarer Sensoren und klinischer Aufzeichnungen, könnte einen großen Fortschritt für die Krankheitsvorhersage und die personalisierte Medizin bedeuten.

Ziel dieses Beitrags ist es, anhand von Beispielen aus unserer Forschung einen Überblick über das Potential und die Herausforderungen der Entwicklung von

KI zur Früherkennung von Krankheiten zu geben. Noch bestehen Hürden für die bestmögliche Nutzung der erforderlichen Daten, vor allem, weil es an Interoperabilität und Standardisierung mangelt. Die Entwicklung von geeigneten Maschinellen Lernverfahren stellt einen entscheidenden Flaschenhals dar.

Krankheitsrisiko

Krankheitsrisiken werden mittels Risikoscores bewertet. Ein Risikoscore ist eine Zahl, deren Höhe das zu erwartende Risiko eines Individuums für das Auftreten einer Krankheit darstellt. Damit können beispielsweise Individuen ermittelt werden, die ein höheres (oder niedrigeres) Krankheitsrisiko haben und Früherkennungsmaßnahmen gezielt in Risikogruppen eingesetzt werden. Für deren Berechnung werden in großen Populationsstudien statistische Zusammenhänge zwischen Krankheitsphänotypen und möglichen Risikofaktoren wie Umwelteinflüssen, Lebensstil und genetischen Mutationen hergestellt. Dazu werden in diesen Populationsstudien statistische Tests durchgeführt, um zu detektieren, ob ein Risikofaktor in Individuen mit einer bestimmten Krankheit signifikant über- oder unterrepräsentiert ist, und zwar im Vergleich zu dem, was durch reinen Zufall zu erwarten ist. In generalisierten linearen Modellen wird der Effekt auf das Krankheitsrisiko geschätzt, also die Veränderung des Risikos, wenn ein bestimmter Faktor vorhanden ist. Alter, Geschlecht, Abstammung, Alkoholkonsum, Übergewicht, Ernährung, Bewegung und Tabakkonsum sind wichtige Faktoren, die mit einer Vielzahl von Erkrankungen wie Herz-Kreislauf-Erkrankungen oder Krebs assoziiert sind. Jeder einzelne assoziierte Faktor kann als Risikofaktor das Krankheitsrisiko erhöhen oder es als protektiver Faktor reduzieren.

In unserer Forschung befassen wir uns vor allem mit dem Einfluss genetischer Mutationen auf Phänotypen und der Modellierung von genetischem Risiko. In den letzten Jahrzehnten haben genomweite Assoziationsstudien in der Forschung eine noch nie dagewesene Menge neuer biologischer Erkenntnisse über menschliche Krankheiten erbracht, darunter Herz-Kreislauf-Erkrankungen, Typ-2-Diabetes, entzündliche Darmerkrankungen, Brust-, Prostata- und Darmkrebs. Es wurde eine Vielzahl genetischer Varianten entdeckt, die mit Krankheitsphänotypen assoziiert sind. Die Schätzwerte der Effekte genetischer Varianten bilden zusammengefasst die Basis für polygene Risikoscores, also Scores, die auf den Effekten mehrerer genetischen Einflussgrößen beruhen. Für einige häufige Krankheiten mehren sich die Hinweise darauf, dass genetische Scores

ein Vorhersagepotenzial haben, das dem von bekannten klinischen Risikofaktoren vergleichbar ist [1].

Zur Verbesserung von Risikoscores gibt es mehrere Ansätze: So könnte man größere und diversere Stichproben wählen, indem zum Beispiel mehrere Biobanken zusammenlegt werden, genauere Messungen von Phänotypen, zum Beispiel mittels bildgebender Verfahren, und eine integrative Analyse mit Gesundheitsdaten aus elektronischen Gesundheitsakten und molekularen Messungen vornehmen. Je höher die Anzahl an Probanden, desto höher ist die Genauigkeit der Studienergebnisse, d.h. hier können Risikoeffekte genauer abgeschätzt sowie neue Effekte von seltenen Varianten und Varianten mit kleinem Einfluss entdeckt werden. Während in Europa bereits einige große nationale und regionale Gesundheitsdatenbanken mit Stichprobengrößen von zehn- bis hunderttausenden Individuen bestehen (z. B. UK Biobank, Estonian Biobank, deCode Genetics, FinnGen, Genomics England und die NaKo Gesundheitsstudie), verspricht deren Zusammenführung die Möglichkeit, mehrere Millionen von Individuen gemeinsam analysieren zu können und damit sowohl die Stichprobengröße als auch die Diversität der Studienpopulation zu erhöhen. Letzteres adressiert das Problem, dass einzelne Bevölkerungsgruppen wie beispielsweise weiße Europäer in einzelnen Studien überrepräsentiert sind. Ein derartiger Mangel an Diversität hat zur Folge, dass die Risikoscores ungenauer im Hinblick auf unterrepräsentierte Bevölkerungsgruppen sind. Es bestehen jedoch regulatorische und praktische Hürden, die neben einem Mangel an Interoperabilität die Zusammenführung von Gesundheitsdaten behindern. Föderierte Lernverfahren und statistische Metaanalysen verfolgen daher den alternativen Ansatz, die Daten getrennt zu analysieren, ohne dass die Daten dafür bewegt werden müssen, und nur die Modellparameter zusammenzuführen. Mit dem Anfang 2021 ins Leben gerufene INTERVENE-Projekt verfolgen wir das Ziel, die auf den Gesundheitsdaten von fast zwei Millionen Europäer:innen und Amerikaner:innen basierenden Analyseergebnisse – in einer föderierten Analyseplattform zusammenzuführen und auf dieser paneuropäischen Stichprobe integrierte Risikoscores zu berechnen. Das heißt, diese Risikoscores berücksichtigen sowohl die genetischen Faktoren als auch Informationen aus anderen Gesundheitsdaten wie beispielsweise elektronische Gesundheitsakten und molekulare Messungen.

1 Mars N, Koskela JT, Ripatti P, Kiiskinen T, Havulinna AS, Lindbohm, JV, ... & Ripatti S. 2020. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nature medicine, 26(4): 549557; Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, & Antoniou AC. 2019. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genetics in Medicine, 21(8): 1708–1718.

Bildgebende Verfahren zur Früherkennung

Bildgebende Verfahren erlauben es, nichtinvasiv quantitative Hinweise auf das Entstehen von Krankheiten zu messen. Sie werden daher zunehmend für die Früherkennung von Krankheiten eingesetzt. Bisher obliegt die komplexe Analyse und Interpretation von medizinischen Bildern dem Fachpersonal und ist daher stets subjektiv. Entscheidungen von Analysten variieren mit der jeweiligen Erfahrung und Berufsausbildung. Auch situationsbedingte Faktoren wie Ablenkung, Konzentration und Müdigkeit beeinflussen die Analyse. KI-basierte CAD-Verfahren versprechen eine objektive Analyse medizinischer Bilder. Bildanalyseverfahren basieren zunehmend auf künstlichen Neuronalen Netzen, deren Modellparameter mittels Überwachten Lernverfahren trainiert werden. Den Neuronalen Netzen werden im Zuge des Trainings iterativ eine große Anzahl an Bildern sowie eine gewünschte Zielgröße, wie beispielsweise die Klassen „gesund" oder „krank", präsentiert. Andere mögliche Zielgrößen können Zahlen wie ein Organvolumen, Klassen wie gutartiger oder bösartiger Tumorschnitt, aber auch strukturierte Objekte wie die Lokalisierung oder die Umrandung eines Tumors sein. Ziel des Trainings ist es, eine Parameterkonfiguration der Neuronen zu finden, so dass automatisch Muster im Bild identifiziert werden, die es erlauben, die gewünschte Zielgröße akkurat vorherzusagen. In jedem Trainingsschritt berechnet das Neuronale Netz basierend auf einem Bild und seiner momentanen Parameterkonfiguration eine Vorhersage der Zielgröße. Mittels des Backpropagationalgorithmus wird effizient ein Parameterupdate berechnet, das den Unterschied zwischen der Vorhersage und der korrekten Zielgröße schrittweise verringert. Diese Schritte werden so lange auf verschiedenen Bildern wiederholt, bis die Vorhersagen und die Zielgrößen ähnlich genug sind, bis also beispielsweise eine gewünschte Fehlerrate auf den Trainingsdaten unterschritten wird.

In der Vukuzazi-Gesundheitsstudie des Africa Health Research Institute in Durban wurden in der südafrikanischen Provinz KwaZulu-Natal in ländlichen Gemeinden mit extrem hoher Prävalenz für Tuberkulose und HIV zur Erkennung der Tuberkulose Röntgenbilder des Brustkorbs der Studienteilnehmer:innen aufgenommen. Ziel war es, basierend auf Anzeichen von Tuberkulose in den Röntgenbildern, die Studienteilnehmer:innen für einen mikrobiologischen Test für Tuberkulose zu triagieren, um Budget einzusparen. Da diese Röntgenbilder unabhängig von vorherigen Symptomen oder einem Verdacht auf Tuberkulose von allen Studienteilnehmer:innen aufgenommen wurden, ähnelt diese Studie dem Einsatz von Röntgen zur Früherkennung von Tuberkulose. Im Rahmen der Studie haben wir einen KI-Algorithmus zur Detektion von Tuberkulose

(CAD4TB [2]) mit den Einschätzungen eines erfahrenen Radiologen verglichen.[3] Der Algorithmus und der Radiologe hatten eine vergleichbare Vorhersagequalität im Hinblick auf Sensitivität und Spezifizität. Das heißt, der KI-Algorithmus und das medizinische Fachpersonal gelangen zu vergleichbaren Analyseergebnissen. Jedoch haben wir basierend auf der ersten Vukuzazi-Pilotstudie, in der sowohl das Röntgen als auch ein mikrobiologischer Test eingesetzt wurden, gesehen, dass der Algorithmus eine deutlich niedrigere Sensitivität im Vergleich zu einer vorhergehenden Evaluation desselben Algorithmus in einer anderen Studie zeigte, die in einer Tuberkuloseklinik in Tansania erstellt wurde.[4] Um in Vukuzazi eine annehmbare Sensitivität von über 80 % zu erreichen, musste der Tuberkulosedetektionsgrenzwert stark gegenüber der Herstellerempfehlung heruntergesetzt werden, so dass nur eine Spezifizität von 66 % in der Triage möglich war. Das heißt, um einen Großteil (80 %) der vorhandenen Tuberkulosefälle zu finden, mussten so viele mikrobiologische Tests durchgeführt werden, dass ein Drittel der Tests negativ war. Dieses Ergebnis steht in Einklang mit einer kürzlich durchgeführten bevölkerungsbasierten Untersuchung in Uganda, die gezeigt hat, dass Personen, bei denen im Rahmen gemeindebasierter Untersuchungen eine aktive Tuberkulose diagnostiziert wurde, weniger Symptome aufweisen als Personen, die in Gesundheitseinrichtungen diagnostiziert wurden. Dieser Unterschied lässt sich dadurch begründen, dass in einer Klinik der Nachweis von Tuberkulose meist nach einem Verdacht basierend auf der Beobachtung vorheriger Symptome durchgeführt wird, wohingegen in einer gemeindebasierten Studie auch scheinbar gesunde Studienteilnehmer:innen untersucht und somit (noch) asymptomatische Tuberkulosefälle gefunden werden.

Die Diskrepanz zwischen den Ergebnissen von Untersuchungen innerhalb und außerhalb von Klinken verdeutlicht, dass klinische Patientendaten sich nur begrenzt für die Entwicklung und Evaluation von KI-Modellen für die Früherkennung von Krankheiten eignen.

2 Murphy K, Habib SS, Zaidi SMA, Khowaja S, Khan A, Melendez J, et al. & van Ginneken, B. 2020. Computer aided detection of tuberculosis on chest radiographs: an evaluation of the CAD4TB v6 system. Scientific reports, 10(1): 111.

3 Fehr J, Konigorski S, Olivier S, Gunda R, Surujdeen A, Gareta D, ..., Lippert C & Wong EB. 2021. Computer-aided interpretation of chest radiography reveals the spectrum of tuberculosis in rural South Africa. NPJ digital medicine, 4(1): 110.

4 Breuninger M, van Ginneken B, Philipsen RH., Mhimbira F, Hella JJ, Lwilla F. … & Reither K. 2014. Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa. PloS one, 9(9), e106381.

Eine Alternative für das Training von Modellen für die Früherkennung stellen prospektive Bevölkerungsstudien dar. Im Gegensatz zu klinischen Patientenkohorten wird in Bevölkerungsstudien ein Querschnitt durch die gesamte Bevölkerung gewählt. Die Proband:innen werden oft prospektiv, also über einen gewissen Zeitraum beobachtet, so dass man das Entstehen von Krankheiten feststellen kann. Wenn anfangs gesunde Proband:innen eine Krankheit entwickeln, lässt sich in den vorausgegangenen Messungen nach Anzeichen auf das Entstehen der Krankheit sowie nach möglichen Risikofaktoren suchen.

Mittels Magnetresonanztomographie wurden seit dem Start der UK-Biobankstudie bis heute 3D-volumetrische Bilder des gesamten Körpers von 45.000 Individuen (ca. 20 % der in der UK Biobank erfassten Individuen) aufgenommen, wobei die Zielgröße bei 100.000 Individuen liegt. Während diese Bilder einen detaillierten Einblick in die Phänotypen dieser Individuen geben und ein wahrer Schatz für die Entwicklung von KI-Modellen für die Früherkennung von Krankheiten sind, stellen sie neue Anforderungen an die Analysemethoden. Die Sichtung und Analyse dieser 3D-volumentrischen Bilddaten durch medizinisches Fachpersonal wie beispielsweise Radiolog:innen sind zeitaufwendig. Die Verwendung von überwachten maschinellen Lernverfahren wie künstliche Neuronale Netze sind ein Ansatz, um die Daten automatisch zu analysieren. Neuronale Netze können in den Bildern automatisch Kennzahlen wie Volumina quantifizieren, Organe markieren oder das Auftreten von Objekten wie Tumore detektieren. Bevor sich ein Neuronales Netz jedoch auf den gesamten Datensatz anwenden lässt, muss es trainiert werden. Für das Training eines Neuronalen Netzes für eine dieser Anwendungen sind Überwachungssignale in Form von Beispielannotationen der gewünschten Zielgröße erforderlich. Diese Datenannotationen müssen durch medizinisches Fachpersonal weiterhin auf der Grundlage einer hinreichend großen Untermenge an Daten manuell erstellt werden. Wenn auch die Anzahl benötigter Annotationen oft signifikant kleiner ist als die Gesamtheit der Daten, stellt deren Erstellung einen Flaschenhals für die Entwicklung von überwachten neuronalen Netzen dar. Auch für die Evaluation der Vorhersagegenauigkeit des neuronalen Netzes werden weitere manuell erstellte Annotationen benötigt. Sichtung und Korrektur von generierten Modellvorhersagen benötigen also weitere manuelle Arbeitsschritte. Da dieser klassische Entwicklungsprozess für jede einzelne Kenngröße durchgeführt werden muss, ist er nur begrenzt zu großen Datensätzen wie der UK Biobank skalierbar.

Um solche großen Datensätze vollständig für die Entwicklung von KI-Modellen zugänglich zu machen, erforschen wir daher verschiedene Wege, um den Trainingsprozess Neuronaler Netze effizienter zu gestalten und den Bedarf an manuell

erstellten Bildannotationen zu reduzieren. Mit der Software VISIAN versuchen wir, den Ansatz des Expert-In-the-Loop-Active-Learning zu implementieren. Ziel der sich in der Entwicklung befindlichen Software ist, das Fachpersonal darin zu unterstützen, mit Hilfe eines Neuronalen Netzes effizient alle 3D-volumetrischen Bilder einer großen Kohortenstudie wie der UK Biobank zu annotieren. VISIAN stellt hierfür eine graphische Benutzerumgebung und Annotationswerkzeuge bereit. Diese werden im Hintergrund dazu verwendet, ein Neuronales Netz zu trainieren und damit Annotationen für den gesamten Datensatz automatisch vorzugenerieren, die in derselben Nutzeroberfläche visualisiert und effizient manuell korrigiert werden können. Das Fachpersonal muss hier nicht mehr der Reihe nach alle Datensätze durchgehen, sondern das Modell lässt sich bevorzugt die Stellen, an denen es bei der Analysequalität unsicher ist, durch das Fachpersonal bestätigen oder korrigieren. Diese Korrekturen werden dann iterativ zur Verfeinerung des Neuronalen Netzes eingesetzt und die automatischen Vorhersagen werden neu generiert. Durch die Zusammenarbeit zwischen KI und Experten kann so eine akkurate Segmentierung bei dem gesamten Datensatz erstellt werden.

Darüber hinaus erforschen wir alternative Ansätze zur Reduktion der benötigten Annotationen auf der Basis von Transferlernen und Selbstüberwachtem und Schwachüberwachtem Lernen. Transferlernverfahren nutzen einen anderen, idealerweise größeren „Ursprungsdatensatz", für den Annotationen bereits vorhanden sind, um das Modell zu trainieren. Diese im Transfer erlernten Modelle werden dann in einem Feinabstimmungsschritt an den eigentlichen Zieldatensatz angepasst. Eine solche Feinabstimmung benötigt oft deutlich weniger annotierte Beispiele auf dem Zieldatensatz. Selbstüberwachte Lernverfahren ersetzen den Bedarf an Annotationen auf dem Ursprungsdatensatz, indem sie basierend auf automatisch generierten Zielvariablen ein Neuronales Netz eine Aufgabe lösen lassen. Beispielsweise soll das Neuronale Netz aus volumetrischen Daten automatisch generierte dreidimensionale Puzzles lösen können, was ein Verständnis über die räumliche Anordnung des Bildinhaltes erfordert.[5] In diesem Fall ist die Anordnung der Bildteile die Zielgröße des Neuronalen Netzes. Um dieses Problem zu lösen, muss das Neuronale Netz eine informative Repräsentation des Bildinhaltes lernen. Da die Zielgröße automatisch erstellt werden kann, benötigt sie keine manuell erstellten Annotationen auf dem Ursprungsdatensatz, was die Nutzung von größeren Ursprungsdatensätzen, beispielsweise der UK Biobank, ermöglicht.

5 Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, & Lippert C. 2020. 3d self-supervised methods for medical imaging. Advances in Neural Information Processing Systems, 33: 18158–18172.

Das Neuronale Netz, das diese Repräsentation berechnet, der sogenannte Encoder, kann dann wie im Transferlernen als Grundlage für das Training von überwachten Neuronalen Netzen auf dem Zieldatensatz verwendet werden, so dass der Lernfortschritt des Modells beschleunigt wird, also weniger annotierte Daten benötigt werden. Anhand von Daten aus der UK Biobank haben wir auch gezeigt, dass die Zuordnung von Bilddaten und Genomsequenzierdaten eine geeignete Aufgabe darstellt, um gute Bildencoder für das Transferlernen zu erhalten.[6] Das Neuronale Netz soll hier Bild- und Genomencoder lernen, so dass einerseits die Repräsentationen des Bildes und des Genoms desselben Individuums möglichst ähnlich sind, und andererseits die Repräsentationen des Bildes von einem Individuum und des Genoms eines anderen Individuums möglichst unterschiedlich ausfallen. Da hier eine weitere Datenquelle, also Genomdaten verwendet werden, nennt man diese Verfahren auch schwach überwacht.

Ausblick

Momentan spielt KI-basierte Früherkennung noch eine untergeordnete Rolle in einem Gesundheitssystem, das auf die Behandlung von Krankheiten statt auf die Erhaltung von Gesundheit ausgerichtet ist. Eines der Ziele des INTERVENE-Projektes ist es, gezielte Interventions- und Präventionsprogramme basierend auf integrierten Risikoscores zu entwickeln, zu pilotieren und deren Vorteil zu quantifizieren. In der „UK Biobank Imaging Study" wurde bereits die Prävalenz von Zufallsbefunden der MRT-Untersuchung des gesamten Körpers auf einen Anteil von 12,8 % errechnet; 3,9 % der Befunde sind von ernsthafter Schwere.[7] Dies zeigt das Potential der Magnetresonanztherapie für die Früherkennung auf. Es gibt auch schon erste kommerzielle Ansätze, eine KI-unterstützte Früherkennung von Krankheiten zu etablieren. Beispielsweise hat die amerikanische Firma Human Longevity, Inc. mit Sitz in San Diego, bei der ich von 2015 bis 2017 angestellt war, mit dem Health Nucleus ein Produkt zur Früherkennung entwickelt. Dabei wurden Genomsequenzierung, Mikrobiomsequenzierung und MRT für die Früherkennung eingesetzt. Diese Daten werden mit KI-Algorithmen analysiert und das Ergebnis durch Ärzt:innen aufbereitet. Bei ihrer Einführung im Jahr 2015 hat die Health-Nucleus-Unter-

6 Taleb A., Kirchler M., Monti R., & Lippert C. 2022. ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with Genetics. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

7 Gibson LM, Paul L, Chappell FM, Macleod M, Whiteley WN, Salman RAS, ... & Sudlow CL. 2018. Potentially serious incidental findings on brain and body magnetic resonance imaging of apparently asymptomatic adults: systematic review and meta-analysis. bmj, 363.

suchung 25.000 US-Dollar gekostet und war somit nur einer kleinen Gruppe von wohlhabenden Erstbenutzern zugänglich. Nach Aussage der Firma wurden bei 8 % der ersten 209 Kund:innen „altersbedingte Krankheiten festgestellt, die eine sofortige (< 30 Tage) medizinische Behandlung erfordern", und bei 2 % wurde durch MRT-Untersuchungen Krebs im Frühstadium entdeckt[8]. 2019 wurde der Preis für den Health Nucleus auf 5.500 US-Dollar reduziert. Auch wenn der Nutzen der einzelnen Untersuchungen noch unklar ist, wird das Produkt weiterentwickelt und verbessert. Durch technologische Fortschritte und durch Skaleneffekte, die durch eine breitere Anwendung des Dienstes erzielt werden, aber auch durch den Fokus auf Komponenten mit hohem Kosten-Nutzen-Verhältnis, soll das Produkt zum einen genauer und außerdem erschwinglich für alle werden.

KI-Methoden, kombiniert mit großen elektronischen Gesundheitsdatenbanken, könnten es ermöglichen, pathologische Prozesse in einem frühen Stadium zu erkennen und somit eine Behandlung zu initiieren, bevor bereits größere irreversible Schäden eingetreten sind. Durch gezielt eingesetzte Vorsorgemaßnahmen und frühes Eingreifen könnten wir so vermeiden, dass Menschen, die an komplexen oder seltenen Krankheiten leiden, in großem Umfang überbehandelt oder falsch behandelt werden. Dies stellt die Möglichkeit einer enormen Entlastung für Patient:innen und ihre Familien, Kliniker:innen, Versicher:innen und die Gesellschaft insgesamt in Aussicht.

8 Perkins BA, Caskey CT, Brar P, Dec E, Karow DS, Kahn AM, ... & Venter JC. 2018. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. Proceedings of the National Academy of Sciences, 115(14): 3686–3691.

GUTE DATEN FÜR EINE VERTRAUENSWÜRDIGE KI IN DER MEDIZIN

Tobias Schäffter, Daniel Schwabe, Stefan Haufe

Jeden Tag wird im Gesundheitssystem eine Vielzahl von Daten erhoben, die Informationen über den Gesundheitszustand von Menschen liefern, um gegebenenfalls notwendige Behandlungsschritte frühzeitig einzuleiten. Die Auswertung erfolgt derzeit durch Ärzt:innen entlang medizinischer Richtlinien. Allerdings wird das Potenzial der Daten für die Gesundheitsversorgung bisher nur im geringen Maße genutzt. Durch die schnell voranschreitende Digitalisierung stehen immer mehr Daten für solche Entscheidungen zur Verfügung. Es werden auch zunehmend Methoden der Künstlichen Intelligenz (KI) in der Medizin genutzt, um Zusammenhänge in Daten zu erfassen und Diagnosen mit zuvor gelernten Mustern zu vergleichen. KI-Methoden werden in der Forschung eingesetzt, um bisher Unbekanntes über Krankheiten zu lernen und auf dieser Basis neue Diagnose- und Behandlungsansätze zu entwickeln, die noch gezielter auf die einzelne Patientin, den einzelnen Patienten ausgerichtet sind. Beispielsweise können bevölkerungsbezogene Versorgungsdaten seltene Nebenwirkungen von Therapien aufzeigen und helfen, Patientinnen und Patienten individueller und damit besser zu behandeln. Das gilt insbesondere für komplexe Therapien, etwa bei Krebserkrankungen. KI-basierte Systeme könnten in Tausenden von Datensätzen zu vergangenen Fällen Muster erkennen, mögliche Neben- und Wechselwirkungen identifizieren und herausfinden, welche Faktoren ausschlaggebend für einen positiven Therapieansatz sein können. Neben der Verwendung von KI-Methoden in der Forschung halten diese auch zunehmend Einzug in Medizinprodukte und damit in den Arbeitsalltag behandelnder Ärztinnen und Ärzte. Die Zahl der Medizinprodukte mit KI-Anteil hat sich seit in den letzten 5 Jahren erheblich erhöht mit über 240 zugelassenen Produkten in Europa.[1] Die Mehrheit der Produkte unterstützt dabei die Arbeit von Radiologen und Kardiologen. So kann beispielsweise eine KI-unterstützte Diagnose anhand von Mustern in medizinischen Bildern oder EKG-Signalen oft schneller und präziser durchgeführt werden als durch einen einzelnen Menschen allein. Die Ärztin oder der Arzt erhalten wertvolle Hinweise, die sie mit ihrer medizinischen Erfahrung bewerten, um dann letztendlich medizinische Entscheidungen zu fällen.

1 Muehlematter UJ, Daniore P, Vokinger KN. 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. Mar; 3(3):e195e203. doi: 10.1016/S2589-7500(20)30292-2.

Im Gegensatz zu klassischen Algorithmen, bei denen der Rechenweg definiert wird, werden in der KI andere Verfahren eingesetzt, die nicht notwendigerweise zu erwartbaren Ergebnissen führen. So werden bestimmte KI-Systeme zunächst trainiert, um Zusammenhänge und Gesetzmäßigkeiten aus Daten zu lernen und diese dann auf neue Daten anzuwenden. Solche Methoden des maschinellen Lernens (ML) stellen eine Teildisziplin der künstlichen Intelligenz dar und führten in den vergangenen Jahren bereits zu sehr großen Erfolgen. Maschinelle Lernverfahren sind auch Computerprogramme (Algorithmen), deren Rechenvorschriften allerdings nicht von vornherein feststehen, sondern anhand von Trainingsdaten erlernt werden. Mit falschen oder unvollständigen Daten gespeist, können solche Verfahren der KI auch falsche oder verzerrte Ergebnisse liefern. In den meisten Fällen sind Ergebnisse des ML auch nicht nachvollziehbar. Darüber hinaus können KI-basierte Systeme problematische oder sogar diskriminierende Entscheidungen hervorrufen, etwa, wenn Trainingsdaten nur bestimmte Teile der Bevölkerung, wie z. B. junge weiße männliche Personen, repräsentieren. Es können dann unvorhersehbare und „verzerrte" Ergebnisse für andere Gruppen geliefert werden, z. B. über Ältere oder Frauen, die nicht oder nur zu geringem Teil in den Daten repräsentiert waren. Derzeit gibt es relativ wenige große Datensätze, die öffentlich zugänglich sind und bei der Entwicklung von ML-Verfahren genutzt werden können. Neben Datenschutz- und Datensicherheitsfragen zur Wahrung von Persönlichkeitsrechten spielen dabei auch wirtschaftliche Gründe eine Rolle. So werden viele medizinische Daten im Rahmen von Studien erhoben, die durch Firmen finanziert sind. Diese Daten sind dann die Geschäftsgrundlage von KI-Produkten. Daneben werden Daten, die zur Entwicklung neuer Verfahren genutzt werden, auch durch proprietäre Plattformen gesammelt. Beispielsweise hat vor Kurzem ein Team von Apple eine umfangreiche Studie zur automatischen Detektion von Herzrhythmusstörungen in einem renommierten Fachjournal veröffentlicht[2]. Dazu wurde der Herzrhythmus von mehr als 400.000 Teilnehmern über acht Monate mit einem Algorithmus der Apple Watch überwacht. Bei ca. 0.5 % der Teilnehmer konnte ein unregelmäßiger Herzschlag entdeckt werden. Diese enorme Datenmenge wurde innerhalb kurzer Zeit im Apple-Konzern nach freiwilliger Einwilligung der Nutzer gesammelt, eine Größe, wie sie bei öffentlich finanzierten Studien nur sehr selten erreicht wird. Die Daten wurden von Apple auch genutzt, um die EKG-App bei der zuständigen FDA-Behörde in den USA zuzulassen.

2 Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, Hung G, Lee J, Kowey P, Talati N, Nag D, Gummidipundi SE, Beatty A, Hills MT, Desai S, Granger CB, Desai M, Turakhia MP. 2019. Apple Heart Study Investigators. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. N Engl J Med. 14; 381(20): 1909–1917.

Zusätzlich wurde diese Funktion in einer klinischen Validierungsstudie an 600 Probanden getestet, wobei die Hälfte diagnostiziertes Vorhofflimmern besaß. In der Zulassung wurde klargestellt, dass das aufgenommene EKG der Apple Watch nur informieren, aber nicht diagnostizieren kann. In Europa wurde die EKG-App von Apple separat als Medizinprodukt der Risikoklasse I nach der europäischen Medizinprodukteverordnung (MDR)[3] zugelassen. Interessanterweise wurde nicht die gesamte Uhr als Medizinprodukt zertifiziert, sondern nur die Software, wobei dafür eine Selbstkonformitätsbewertung ausreichte. Das Beispiel zeigt, dass derzeit eine unabhängige Prüfung des Algorithmus und der zu Grunde liegenden Daten fehlt. Die eigentliche Zulassung erfolgte über eine klinische Studie, was im Allgemeinen recht aufwendig ist und derzeit vor allem von kleinen und mittelständischen Unternehmen als Innovationshemmnis gesehen wird. Klinische Studien bewerten im Grunde nur die Ergebnisse in einem engen Einsatzbereich der Studienkohorte und erlauben nur bedingt Aussagen über die Qualität für einen breiteren Einsatz. Es wird seit einigen Jahren an Verfahren gearbeitet, welche das Verhalten von KI-Methoden charakterisieren. Dennoch fehlen allgemein gültige Qualitätsregeln und Prüfverfahren. Um die Verwendung von KI in Medizinprodukten sicher und verlässlich zu machen und um deren Vertrauen und Akzeptanz in der Gesellschaft zu schaffen, sind neue Ansätze für eine digitale Qualitätsinfrastruktur von Nöten.

Qualität und Prüfverfahren

Die Qualität der Trainings- und Testdaten spielt eine fundamentale Rolle bei der Qualitätssicherung aller ML-Anwendungen. Wie wichtig eine gesicherte Datenlage für eine solche KI-Anwendung ist, hängt von der Kritikalität des Algorithmus ab. Der Entwurf des Artificial Intelligence Act der EU[4] etabliert hierzu vier Risikogruppen und damit einhergehende Prüfanforderungen für ML-Anwendungen. Dabei zählen Medizinprodukte und In-Vitro-Diagnostika zu den Hochrisikoanwendungen [AI Act, Punkt (30)]. Für diese gilt die Forderung nach einer „hohen Qualität der Datensätze, die in das System eingespeist werden, um Risiken und diskriminierende Ergebnisse so gering wie möglich zu halten".[5]

3 EU-Medical Device Regulation, MDR 2017/745.

4 EU AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

5 EU AI Act Richtlinien: https://ec.europa.eu/commission/presscorner/detail/de/ip_21_1682

Derzeit fehlen allerdings klar definierte Vorgehensweisen, wie solche allgemeinen Anforderungen in der Praxis umgesetzt werden können. So beklagen laut einer Umfrage der Unternehmensberatung Price Waterhouse Coopers[6] zwanzig weltweit führende Medizinproduktehersteller die „Überregulierung" als die größte Bedrohung für Innovationen. Kürzlich stellte die USA eine offizielle Anfrage bei der Welthandelsorganisation, in der die neue Medizinprodukteverordnung der EU als großes Handelshemmnis für medizinische Geräte aus den USA zum EU-Markt beschrieben wird[7], insbesondere für Geräte, die einen hohen Softwareanteil aufweisen. Es sind also klare Leitlinien und Verfahrensanweisungen für Medizinprodukte mit KI-Komponenten notwendig, um die Qualität der Daten und des KI-Verfahrens zu prüfen. Da bei der Formulierung der europäischen Medizinprodukteverordnung nur im geringen Maße an neue ML-Verfahren gedacht wurde, wird eine Anpassung erwartet. Dazu müssen Test- und Prüfverfahren für KI-Verfahren entwickelt werden, deren Nutzen wiederum in der Anwendung demonstriert werden muss, um Akzeptanz in der Gesellschaft zu schaffen.

Datenqualität

Das Grundprinzip des ML ist das Erlernen von Zusammenhängen in Daten. Aufbauend auf den gelernten inhärenten Strukturen der Daten trifft ein Algorithmus dann Vorhersagen und Entscheidungen. Die Qualität der Trainingsdaten, die zum Lernen verwendet werden, hat daher einen entscheidenden Einfluss auf die Funktionsweise und Qualität einer ML-Anwendung. Es gibt derzeit eine intensive Diskussion darüber, dass der Einsatz von ML zum Teil zu falschen oder diskriminierenden Entscheidungen führen kann. Dies liegt weniger am Algorithmus als an fehlerhaften oder unvollständigen Trainingsdaten, die die Qualität beeinflussen. Aus diesem Grunde ist eine Prüfung und Absicherung der Datenqualität unerlässlich. Beispielsweise sieht die zuständige Behörde in den USA die Wahl der Trainings- und Testdaten als Schlüsselkomponente für den erfolgreichen Einsatz von ML-Techniken. Daten sollten daher unter Verwendung von Qualitätssicherungs- und -Managementsystemen erhoben werden.[8] Dies beinhaltet u. a. Protokolle zur Datenerhebung, Bestimmung eines Referenzstandards sowie die Auditierung von

6 Price Waterhouse Coopers; 20th CEO Survey: Healthcare industry key findings 2017.

7 World Trade Organization (G/TBT/W/679); Statement by the USA to committee on technical barriers to trade; Juli 2019 link: G/TBT/W/679 24 July 2019 (19-4907) Page

8 FDA "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning [AI/ML] Based Software as a Medical Device [SaMD]"

Test- und Trainingsdaten innerhalb der Hersteller-Organisation. Allerdings ist derzeit keine unabhängige Prüfung notwendig. Auch in Europa wird im Rahmen des „Artificial Intelligence Act“ [9] eine Ausgestaltung von Verfahren zur Datenqualität für KI-Verfahren in der Medizin gefordert. Dabei stehen die Qualitätsmerkmale wie Richtigkeit, Präzision, Vergleichbarkeit, Vollständigkeit und Repräsentativität von Daten im Vordergrund.

Damit eine ML-Anwendung die „richtigen“ Entscheidungen treffen kann, müssen die Trainingsdaten eine hohe Genauigkeit aufweisen. Die Genauigkeit ist eine zentrale Fragestellung der Metrologie, der „Wissenschaft des Messens“, wobei hier zwischen Richtigkeit und Präzision unterschieden wird. Die Richtigkeit (oder Fehler, „Bias“) beschreibt die Nähe (oder Ferne) eines Datenwertes zu einem „wahren“ Wert (meist einem Referenzwert). Dagegen beschreibt die Präzision die Abweichung der Daten untereinander, d. h. wieweit die Datenwerte um einen Mittelwert streuen. Systematische Abweichungen können entstehen, wenn beispielsweise unterschiedliche Messsysteme verwendet werden, deren Messungen im Mittel streuen und eine unterschiedliche Nähe zum wahren Wert haben. Beide Qualitätsmerkmale haben eine große Bedeutung für die Entwicklung von ML-Verfahren. Insbesondere führt ein großer systematischer Fehler („Bias“) zu einem verzerrten Lernergebnis, so dass in der Anwendung des Erlernten, neue Werte konsistent falsch eingeordnet werden. Üblicherweise soll aber eine Abweichung von einem Referenzwert zur Diagnose einer krankhaften Veränderung genutzt werden. Daher ist es wichtig zwischen der systematischen Abweichung aufgrund des Messsystems und der Abweichung vom „Normwert“ aufgrund einer Krankheit zu unterscheiden. Das gilt auch für die Präzision; auch hier muss zwischen einer Streuung der Messungen und der biologischen Varianz unterschieden werden. Der Präzision kommt auch eine wichtige Rolle zu, wenn es darum geht, zuverlässige Entscheidungsschwellwerte festzulegen. Oft werden Daten aus unterschiedlichen Quellen und unter Verwendung unterschiedlicher Messverfahren zusammengeführt. Dabei kommt es zu einer Mischung der verschiedenen Präzisions- und Genauigkeitswerte, welche selbst ein Muster bilden können. Falls diese Herkunft nicht berücksichtigt wird, kann dieses „Muster“ ungewollt erlernt werden und es kann die eigentlichen Merkmale, jene die mit der eigentlichen Krankheit zu tun haben, überdecken. Dies ist vor allem der Fall, wenn sogenanntes „confounding“ vorliegt, z. B. wenn sich die relative Häufigkeit von Diagnosen zwischen Datenquellen unterscheidet. Auf Testdaten angewandt, würde das trainierte ML-Verfahren dann, anstatt die gewünschte Klassifikation der Krankheit auf Basis klinisch-

9 https://artificialintelligenceact.eu, 2021.

relevanter Dateneigenschaften durchzuführen, vorrangig die geschätzte Herkunft (d. h. das Messverfahren) verwenden. Neben der Genauigkeit der Daten aufgrund der ursprünglichen Messung, ist die „ungenaue" Zuordnung der Daten (sog. „Label") durch Experten eine weitere Fehlerquelle beim ML.

Die Vergleichbarkeit von Daten beruht auf der Verwendung eines einheitlichen Referenzsystems. Die beste Möglichkeit, die Vergleichbarkeit von Daten – unabhängig davon, wann und wo sie ermittelt wurden – zu gewährleisten, ist die metrologische Rückführung auf ein gemeinsames Referenzsystem. Mit dem Internationalen Einheitensystem (SI) gibt es ein solches System. Es bildet seit der Meterkonvention von 1875 auch die weltweite Grundlage für Handel in über neunzig Staaten. Das SI-System ist fester Bestandteil in vielen Bereichen des täglichen Lebens. Allerdings ist das SI-System im Gesundheitswesen noch nicht vollständig verbreitet. Es sind große EU-weite Initiativen vonnöten, um die Vergleichbarkeit von medizinischen Messverfahren durch metrologische Rückführung weiter zu verbessern und so auch die Qualität von multizentrischen Studiendaten zu erhöhen.

Repräsentativität und Vollständigkeit sind weitere statistische Qualitätsmerkmale für Daten mit hoher Bedeutung für das ML. Dies umfasst zum Beispiel das Verhindern von Diskriminierung jeglicher Form durch eine statistische Unterrepräsentation bestimmter Datengruppen innerhalb der Trainingsdaten. Damit wird gewährleistet, dass ein ML-Algorithmus lernen kann, richtige Entscheidungen und Vorhersagen für alle vorgesehenen Anwendungsbereiche zu treffen. Neben der Qualität der Trainingsdaten spielt die Qualität der Testdaten ebenfalls eine wesentliche Rolle. Diese werden verwendet, um neue ML-Verfahren sowohl zu testen als auch zu verbessern. Dabei gelten die gleichen Qualitätskriterien wie für Trainingsdaten, allerdings mit höherer Bedeutung einzelner Kriterien. Beispielsweise ist die Repräsentativität der Testdaten entscheidend dafür, die Eignung einer ML-Anwendung für unterschiedliche Szenarien gewährleisten zu können. Wesentlich ist hierbei auch, die Trainingsdaten streng von den Test- und Validierungsdaten zu trennen.

Qualität des ML-Verfahrens

ML-Verfahren werden anhand von Daten trainiert, wobei die Entscheidungsregeln vom System selbst erlernt werden, ohne dass diese Zusammenhänge direkt sichtbar sind. Daher wird häufig von einer „Black Box" gesprochen. Oft wird der Wunsch nach Veröffentlichung und Transparenz der Verfahren geäußert. Dies

greift allerdings recht kurz, da selbst bei der Veröffentlichung aller Werte (sog. Gewichte) eines neuronalen Netzwerkes das Verhalten zwar reproduziert werden kann, dieses oft aber selbst vom Entwickler nicht vollständig verstanden wird. Insgesamt ist der Einfluss der Trainingsdaten auf das Verhalten im maschinellen Lernen so stark, dass eine komplett unabhängige Prüfung der beiden Einzelaspekte nicht sinnvoll und möglich ist. Derzeit werden Kriterien und Metriken für die Beurteilung der Qualität von ML-Verfahren zur Validierung entwickelt. Gerade da medizinische Entscheidungen kritisch sein können, sollte es klar definierte Kriterien geben. Dabei stehen die Qualitätsmerkmale wie Leistungsfähigkeit, kalibrierte Unsicherheitsquantifizierung, Erklärbarkeit, Generalisierbarkeit und Robustheit im Vordergrund.

Die Vorhersagegüte (Performance) ist eines der wichtigsten Kriterien zur Beurteilung von ML-Verfahren. Je nachdem, ob die Zielgrößen kontinuierlich oder kategorial sind, kommen neben klassischen metrologischen Fehlermaßen auch Maße aus der Signaltheorie wie Sensitivität und Spezifizität zum Einsatz. Zur Bestimmung der Vorhersagegüte ist es allerdings notwendig, die „richtigen" Werte (Referenzwerte) zu kennen. Daneben ist aber auch die Unsicherheit von Vorhersagen von ML-Modellen gerade in der Medizin von hoher Bedeutung. Hier geht es darum, dem/der klinischen Entscheidungsträger:in nicht nur einen einzelnen Wert an die Hand zu geben, sondern auch eine Einschätzung, wie sicher diese Vorhersage ist. Wenn die Unsicherheit zu hoch ist, kann ein Arzt oder eine Ärztin dies im Entscheidungsprozess berücksichtigen. Dabei werden zwei Haupttypen der Unsicherheit unterschieden. Die epistemische Unsicherheit beschreibt, was das Modell nicht wissen kann, weil die Trainingsdaten nicht angemessen, unvollständig oder in ihrer Anzahl nicht ausreichend sind oder weil die Komplexität des Modells zur Modellierung der Daten nicht ausreicht. Demgegenüber bezieht sich die aleatorische Unsicherheit auf die inhärente Zufälligkeit der Daten, wobei der Begriff *Aleator* im Lateinischen jemanden beschreibt, der würfelt. Bei genügend Trainingsdaten nimmt die epistemische Unsicherheit ab, während die aleatorische Unsicherheit nicht verringert werden kann, selbst wenn mehr Daten bereitgestellt werden. Insgesamt gibt es für die Bestimmung der Unsicherheit von Entscheidungen verschiedene Ansätze. Klassische Bottom-up-Ansätze der Metrologie werden im GUM-Framework[10] beschrieben. Die grundlegende Idee ist hierbei, dass bekannte Verteilungen und Unsicherheiten der Eingangsgrößen benutzt werden können, um entsprechende Verteilungen und Unsicherheiten der Ausgangsgrößen

10 BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. 2008. Guide to the Expression of Uncertainty in Measurement. JCGM 100:2008, GUM 1995 with minor corrections.

einer Funktion zu schätzen. Dieser Ansatz setzt aber voraus, dass die angewandte Funktion statisch ist und nicht von den Daten selbst abhängt. Dies ist im ML jedoch nicht immer gegeben. Stattdessen kann man das Schätzen eines ML-Modells als inverses Problem auffassen, bei der die Verteilung der Modellparameter aus den beobachteten Daten geschätzt wird. Aus der Verteilung der Parameter kann dann wiederum die Verteilung der Vorhersagen abgeleitet werden. Der Nachteil dieser Modellierung ist jedoch eine hohe Subjektivität, da es notwendig ist, Vorwissen über die Verteilungen der Modellparameter zu spezifizieren. Dieser Nachteil kann bei guter Datenlage jedoch zum Teil durch sogenannte empirische Bayes-Verfahren ausgeglichen werden. Ein weiteres interessantes Paradigma zur Unsicherheitsschätzung ist die Top-down-Modellierung. Hierbei wird beispielsweise die Verteilungsfunktion möglicher Ergebnisse aus den Daten gelernt. Ein Beispiel sind „Monte Carlo Dropout"-Ansätze[11] für Deep Learning, welche die Unsicherheit dadurch bestimmen, dass beim Trainieren einzelne „Neuronen" mit einer gewissen Wahrscheinlichkeit deaktiviert (sog. „Dropout") werden, wodurch unterschiedliche Vorhersagen erzielt werden. Nach verschiedenen Durchläufen und unterschiedlichen Deaktivierungen kann dann eine Verteilung der Ergebnisse und daraus eine Unsicherheit bestimmt werden. Ein weiteres Beispiel sind Methoden, die Unsicherheitsparameter (z. B. Standardabweichung oder Perzentile) direkt schätzen, wie sogenannte „deep ensembles".[12] Dafür werden dem neuronalen Netz weitere Neuronen in der Ausgabeschicht hinzugefügt. Diese Modellierung ist komplett datengetrieben. Dabei werden die gelernten Unsicherheitsintervalle durch geeignete Verlustfunktionen (sogenannte „proper scoring functions") kalibriert, sodass die angegebene Überdeckung zumindest bei der Trainingsstichprobe auch tatsächlich zutrifft. Eine weitere Frage ist, wie Unsicherheitswerte von unterschiedlichen neuronalen Netzen verglichen werden können.

Ziel der „Erklärbarkeit" ist es, die vom ML-Verfahren erlernten Zusammenhänge in den Daten zu kennzeichnen. In den vergangenen Jahren wurden Methoden entwickelt, die den menschlichen Nutzerinnen und Nutzern aufzeigen sollen, wie der ML-Algorithmus zu seiner Entscheidung gekommen ist. Dieser Aspekt ist von zentraler Bedeutung in ML-Anwendungen zur Unterstützung von Medizinern, welche maschinelle Ergebnisse verstehen wollen, um so Vertrauen für eine Entschei-

11 Gal Y, Ghahramani Z. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning; Proceedings of the 33rd International Conference on International Conference on Machine Learning. 48: 1050–1059.

12 Lakshminarayanan B., Pritzel A., & Blundell C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.

dung auf dieser Grundlage zu gewinnen. Das Forschungsgebiet der „erklärbaren KI" (engl. Explainable Artificial Intelligence – XAI) wurde mit großen Forschungsprogrammen in den letzten Jahren sehr stark entwickelt (beispielweise wurden in den USA 70 Millionen Dollar für das „Explainable-AI"-Forschungsprogramm[13] zur Verfügung gestellt). Oft zielt „erklärbare KI" darauf ab, Charakteristika in Daten zu identifizieren, welche einen besonders starken Einfluss auf die getroffene Entscheidung des Verfahrens hatten[14]. Diese Datenbereiche können dann z. B. farbig markiert werden, um die Information für den Benutzer und die Benutzerin sichtbar darzustellen. Eine wichtige Technik ist das „Layer-wise Relevance Propagation" LRP-Verfahren[15], das schrittweise die Entscheidungswerte bestimmt. Allerdings wird bezweifelt, ob der „Einfluss" einer Variablen allein ausreicht, um belastbare Aussagen über das Wirkungsprinzip eines trainierten ML-Modells generell oder auf einem konkreten Datenpunkt zu verstehen. So können sogenannte „Suppressorvariablen" mit Störsignalen beispielsweise einen starken Einfluss auf die Vorhersage haben, ohne selbst irgendeine statistische Abhängigkeit zum Vorhersageziel aufzuweisen[16]. Insgesamt ist festzustellen, dass es im Feld der „erklärbaren" KI derzeit keine ausreichend mathematisch fundierten Definitionen von Korrektheit gibt, mit Hilfe derer die „Erklärgüte" von XAI-Methoden objektiv bewertet werden könnte.

Die Generalisierbarkeit beschreibt die Eigenschaft eines ML-Verfahrens, für möglichst viele verschiedene Eingabedaten und Anwendungsszenarien gültige Ausgaben zu liefern. Dieses Ziel muss während des Trainierens eines Modells berücksichtigt werden, indem eine Überanpassung an die Trainingsdaten verhindert wird. Danach sollten die trainierten Modelle an verschiedenen Daten außerhalb des Trainingsdatensatzes getestet werden, um ihre Generalisierbarkeit bewerten zu können. Beispielsweise kann durch eine gezielte Wahl von Testdaten, die sich in einigen ihrer Eigenschaften von den Trainingsdaten unterscheiden (sog. „out-of-distribution data"), die Generalisierbarkeit einer ML-Anwendung untersucht werden. Während Generalisierbarkeit die Fähigkeit von Modellen ist, Datenpunkte außerhalb der Trainingsdaten vorherzusagen, bezieht sich die Robust-

13 Voosen P. 2017. The AI detectives. Science. 357(6346): 22–27.

14 Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. 2019. Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1): 1–8.

15 Montavon G, Samek W, Müller KR. 2017. Methods for Interpreting and Understanding Deep Neural Networks Digital Signal Processing, 73: 115.

16 Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J D, Blankertz B & Bießmann, F. 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87: 96–110.

heit auf die Stabilität eines ML-Verfahrens gegenüber äußeren und feindlichen Einflüssen. Das Kriterium „Robustheit" wird meist in der Softwaresicherheit verwendet und beschreibt die Eigenschaft eines ML-Systems, gegenüber Angriffen durch falsche Daten zu bestehen. Schon im Jahr 2014 haben Forscher von Google die Anfälligkeit bestimmter neuronaler Netze nachgewiesen und konnten Daten so manipulieren, dass ML-Verfahren falsche Entscheidungen fällten. Solche feindlichen („adverserial") Daten werden im Normalfall vom menschlichen Betrachter nicht erkannt, da sie oft nur auf minimalen Änderungen der Originaldaten basieren. Sie führen aber zu falschen ML-Entscheidungen, da sie zu Angriffszwecken konstruiert wurden. Die entwickelten Verteidigungsmethoden können verwendet werden, um ML-Verfahren, die über ihre Eingabemodalitäten Risiken zu sicherheitsrelevanten Manipulationen des Systemverhaltens bieten, abzusichern und mit geeigneten Verfahren die Robustheit des Systems zu messen. Eine Möglichkeit die Robustheit zu erhöhen, ist die Verwendung von Filtern, welche manipulative Veränderungen reduzieren. Es können auch informationstheoretische Methoden verwendet werden, um zu prüfen, ob Daten weitere potenziell feindliche Informationen enthalten.[17]

Vergleichstests

Neben den genannten Kriterien wird noch an weiteren Eigenschaften gearbeitet, um die Qualität von ML-Verfahren möglichst umfassend bewerten zu können. Solche Kriterien bilden die Grundlage für Vergleiche („Benchmarktests") und erlauben die Definition von Referenzverfahren mit definierter Qualität, an denen sich neue Entwicklungen messen können. Allerdings gibt es aufgrund der Vielfalt von Anwendungen keine allgemein gültige Qualitätseinschätzung, sondern eher Richtlinien. Da die Qualität der ML-Verfahren maßgeblich von der Datenqualität abhängt, ist ein fairer Vergleich nur unter Verwendung gleicher Validierungsdaten möglich. Zur Beurteilung maschineller Lernverfahren werden regelmäßig Wettbewerbe (sog. „Challenges") ausgetragen, in denen die Leistungsfähigkeit der ML-Modelle anhand vorgeschriebener Bewertungskriterien mit standardisierten Datensätzen verglichen werden. In diesem Zusammenhang hat die Definition von Referenzdaten einen hohen Stellenwert. Auch wenn es nicht möglich ist, Referenzdaten für alle Anwendungsfälle zur Verfügung zu stellen, können diese für bestimmte Klassen wie Einzelwertmessungen (z. B. Temperatur, Blutdruck, Sauer-

17 Martin J, Elster C. 2020. Inspecting adversarial examples using the fisher information. Neurocomputing, 382. 80–86.

stoffsättigung), Zeitreihen (Elektrokardiogramm, Pulswellen) oder medizinische Bilder für verschiedene Fragestellungen definiert werden. Leider ist das Angebot an offenen Referenzdatensätze gering bzw. die Größe der Datensätze so klein, dass ein guter Vergleich nicht immer möglich ist. Es gibt aber große internationale Initiativen mit dem Ziel, dies zu verbessern. Beispielsweise wurde kürzlich ein offener Datensatz mit über 21.000 Elektrokardiogramm (EKG)-Messungen von mehr als 18.000 Patienten und Gesunden veröffentlicht.[18] Jede klinische EKG-Messung wurde von 10 Elektroden erfasst und danach von zwei Kardiologen diagnostiziert und entlang 71 standardisierter Klassen eingeordnet. Der Datensatz weist eine gute Verteilung sowohl zu verschiedenen Krankheitsklassen als auch zu Gesunden auf. Darüber hinaus sind auch verschiedene demografische Merkmale (z. B. Alter und Geschlecht) repräsentiert. Neben der diagnostischen Qualität gibt es auch Hinweise zur unterschiedlichen EKG-Signalqualität. Schwerpunkt des Datensatzes ist die Definition von Unterdatensätzen für Training, Test und Validierung mit möglichst einheitlicher Repräsentativität. Dabei wurde insbesondere auf eine hohe Label-Qualität bei den Test- und Validierungsdaten gelegt. Der Datensatz wurde kürzlich für eine Vergleichsstudie verwendet[19], um bestehende ML-Verfahren entlang definierter Kriterien (Leistungsfähigkeit, Unsicherheit, Erklärbarkeit) zu untersuchen und so einen strukturierten Ansatz für zukünftige Vergleiche und eine Einordnung neuer Verfahren zu etablieren. Eine potenzielle Fehlerquelle von Vergleichstests anhand solcher klinischer Datensätze ist, dass Fehler bei der Beurteilung durch Mediziner nicht ausgeschlossen werden können und nur durch erheblichen Aufwand, d. h. Beurteilung der Daten durch möglichst viele Experten, minimiert werden können. Ein alternativer Ansatz ist die Verwendung von synthetischen Daten. Diese simulierten Messdaten werden beispielsweise durch biophysikalische Modelle anhand wohl definierter Parameterwerte erzeugt und erlauben daher eine bessere Zuordnung und exakte Berechnung von Fehlern. Im Rahmen des EU-Projektes Medalcare[20] wurden dazu EKG-Daten einer virtuellen Population erstellt.[21] Solche simulierten Daten erlauben es, sowohl die Präzision als auch die Genauigkeit zu ändern, so dass der Einfluss der Datenqualität auf ein ML-Verfahren untersucht werden kann und auf diese Weise Vorgaben

18 Wagner P, Strodthoff N, Bousseljot R.-D., Kreiseler D, Lunze F, Samek W, Schaeffter T. 2020. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 7: 154.

19 Strodthoff N, Wagner P, Schaeffter T, Samek W. 2020. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL, IEEE J Biomed Health Inform.

20 EU-EMPIR Projekt „Medalcare". 2019. https://www.ptb.de/empir2019/medalcare/home/

21 Nagel C, Schuler S, Dössel O, Loewe A. 2021. A bi-atrial statistical shape model for large-scale in silico studies of human atria: Model development and application to ECG simulations. Med Image Anal.74:102210. doi: 10.1016/j.media.2021.102210.

für ein Mindestmaß an Messdatenqualität festgelegt werden können. Besonders interessant ist es, bewusst bestimmte Daten im Training wegzulassen, zu verändern bzw. eine falsche Zuordnung (Label) der Daten durchzuführen, um so die Generalisierbarkeit und Robustheit eines Verfahrens zu untersuchen. Neben dem genannten EKG-Beispiel gibt es viele weitere Initiativen, insbesondere in der biomedizinischen Bildgebung.

Zusammenfassung

Verfahren der künstlichen Intelligenz beruhen auf Daten und Algorithmen. Für beide sind neue Ansätze in der Qualitätssicherung notwendig, um die Zertifizierung und den Einsatz von KI-Verfahren zu beschleunigen und gleichzeitig ein Vertrauen in der Gesellschaft zu fördern. Da die Qualität der Verfahren wesentlich von der Datenqualität abhängt, braucht es Richtlinien für eine objektive Bewertung von Datensätzen. Dies gilt sowohl für öffentliche als auch für private Daten, bzw. Daten, die aus Geschäftsmodellgründen in Unternehmen verbleiben müssen. Richtlinien zur Datenqualität können auch für eine Kuratierung von klinischen Daten zu Referenzdatensätzen genutzt werden, welche sowohl in der Forschung als auch in der Produktentwicklung eingesetzt werden sollten. Insgesamt brauchen Medizinprodukte mit ML in Zukunft eine zusätzliche ML-bezogene Konformitätsbewertung. Hierzu sollten sowohl Trainingsdaten als auch Testdaten von unabhängigen Stellen bewertet werden. Derzeit werden solche regulatorischen Ansätze von Herstellern oft als Hürde empfunden, mittelfristig können diese aber zu einem Wettbewerbsvorteil insbesondere gegenüber den USA und China führen. Eine Bewertung der Qualität der verwendeten Daten und der entwickelten Verfahren sollte auch Auswirkungen auf die weiterhin notwendigen klinischen Studien haben. So könnten klinische Studien kleiner ausfallen, wenn eine hohe Qualität der Trainingsdaten nachgewiesen wurde, was letztendlich zu einer schnelleren Markteinführung führen kann. Daneben schafft ein Qualitätssiegel auch Vertrauen bei Kundinnen und Kunden und führt somit zu einem potenziellen Wettbewerbsvorteil gegenüber Produkten und Dienstleistungen anderer Anbieter ohne Qualitätssiegel. Das Label „Made in Germany" bzw. „Made in Europe" würde als hoher Qualitätsstandard so auf das digitale Zeitalter übertragen werden, um eine Vertrauensbildung beim Verwender und eine Erhöhung des Wettbewerbsvorteils von Herstellern zu fördern. Es ist ein wirtschafts- und gesellschaftspolitischer Rahmen notwendig, der durch folgende Aktivitäten unterstützt werden sollte:

- Entwicklung von europäischen Zulassungsvorschriften für eine zuverlässige und vertrauenswürdige KI in der Medizin.

- Entwicklung europäischer Standards und Normen für Daten- und KI-Qualität.

- Schaffung sog. Reallabore, um in Ermangelung verbindlicher regulatorischer Vorschriften bereits gemeinsam Richtlinien zwischen Herstellern, Behörden, Benannten Stellen, medizinischen Experten und Patientinnen und Patienten zu erarbeiten. Dies beinhaltet die Definition von notwendigen Qualitätskriterien sowohl für Datensätze für ML-Verfahren als auch für ML-Verfahren.

- Interdisziplinäre Zusammenarbeit von Experten an medizinischen Zentren, um hochqualitative Datensätze zu generieren.

- Bereitstellung von Referenzdaten und Vergleichstests für Forschung und Entwicklung.

- Vertrauenswürdige Prüfung von nicht-öffentlich zugänglichen Daten und ML-Verfahren durch unabhängige Stellen und Vergabe von Qualitätssiegeln.

- Regeln für notwendige klinische Studien zum Nachweis der Wirksamkeit eines ML-Systems.

- Erhöhung der gesellschaftlichen Akzeptanz gegenüber vertrauenswürdigen KI-Verfahren durch qualitätsgesicherte KI.

- Motivation zur Datenspende, um große Datensätze für die KI-Forschung zum Nutzen aller zur Verfügung stellen zu können.

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).

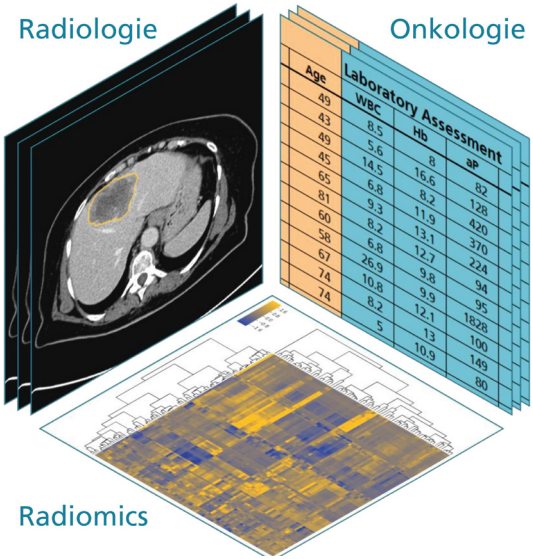


Abb. 1: Radiomics zur Aufdeckung von Zusammenhängen zwischen radiologischen und klinischen Daten

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-

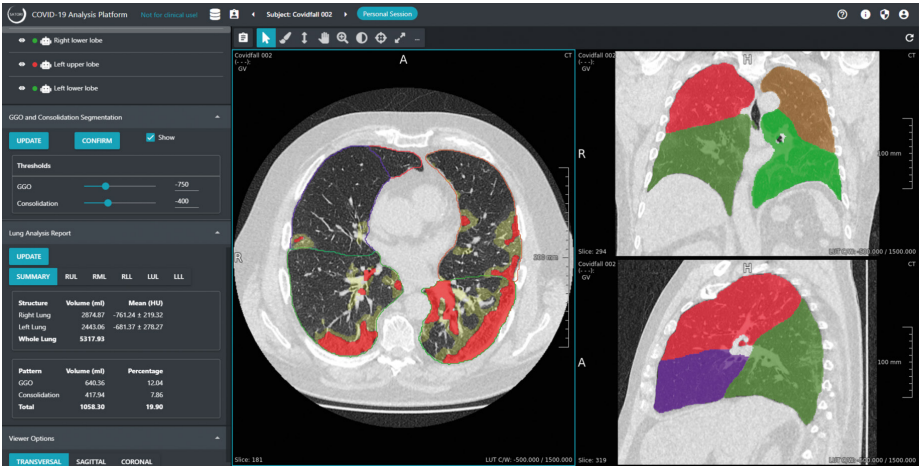


Abb. 2: Prototypische Plattform zur interaktiven Sichtung und Kuratierung von computertomographischen Daten der Lunge

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).

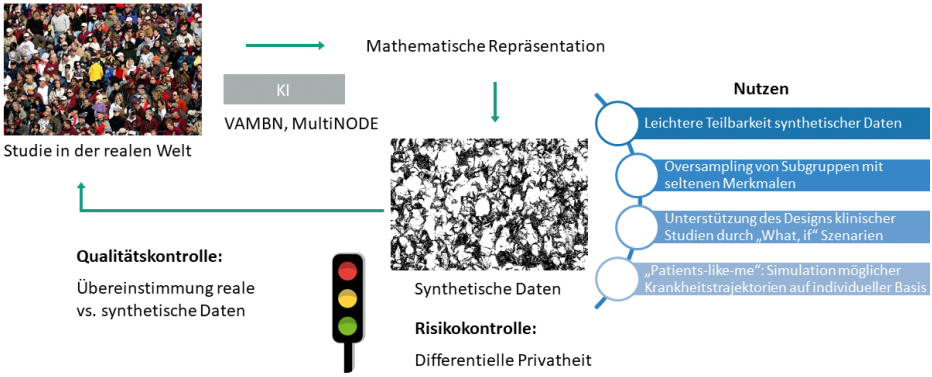


Abb. 3: Erzeugung synthetischer Daten mit Hilfe von KI-Verfahren

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.

NACHHALTIGE MEDIZIN BRAUCHT DIGITALE SOUVERÄNITÄT[1]

Rico Barth, Peter Ganten, Manuela Urban

Zusammenfassung

Es gibt wohl keinen Lebensbereich, in dem die Digitalisierung so viele Chancen und Gefahren birgt wie in der Medizin. Die digitale Vernetzung von Forschung und Entwicklung mit Erkenntnissen und Daten aus der Praxis ermöglicht Fortschritte in der Diagnose und Therapie, die ohne Technologie auf höchstem Niveau undenkbar sind. Information ist eine der wertvollsten Ressourcen unserer Zeit und ihre sinnvolle Nutzung und Verbreitung eröffnet auch in der Zukunft ungeahnte Chancen für die Gesundheit der Menschen weltweit. Gleichzeitig berührt Gesundheit die Intimsphäre jedes Menschen, deshalb ist es gerade hier besonders wichtig, dass jeder Missbrauch dieser persönlichen Informationen in der Nutzung, Sammlung und Weitergabe strikt unterbunden wird. Open Source und die damit verbundene digitale Souveränität ermöglicht ein Zusammenspiel von Offenheit, Transparenz, offenem Austausch, Datensouveränität und Datenschutz. Als Kompetenznetzwerk von Open-Source-Unternehmen und -Experten setzt sich die Open Source Business Alliance seit Jahren für die Förderung von digitaler Nachhaltigkeit, Resilienz und digitaler Souveränität ein und steht als Verband beratend und unterstützend zur Verfügung, um Open-Source-Technologie auch im Bereich der Medizin zum nationalen, europäischen und internationalen Standard zu machen.

Einleitung

Die Wissenschaft lebt seit der Moderne davon, Wissen und Daten zu teilen, um die Qualität der wissenschaftlichen Arbeit durch eine Community-Kontrolle (Peer Review) zu sichern und Exzellenz und Lösungsmöglichkeiten für komplexe Probleme durch breite Zusammenarbeit zu ermöglichen. Die Freiheit und Selbstbestimmtheit wissenschaftlicher Arbeit kann im digitalen Zeitalter nur gesichert werden, wenn auch der Zugang zu wissenschaftlichen Daten und die Verfügbarkeit und Gestaltbarkeit digitaler Werkzeuge frei und offen bleiben. Digitale Souveränität

1 Aus Gründen der besseren Lesbarkeit wird bei Personenbezeichnungen und personenbezogenen Substantiven das Maskulinum verwendet. Gemeint sind im Sinne der Gleichbehandlung grundsätzlich alle Geschlechter. Die verkürzte Sprachform hat redaktionelle Gründe und beinhaltet keine Wertung.

steht für die Fähigkeit, Datenflüsse kontrollieren und digitale Verfahren beeinflussen und weiter entwickeln zu können. Für die medizinische Praxis ist beides fundamental, um höchste Qualität, Verfügbarkeit und schnelle Innovation digital bestimmter medizinischer Leistung zu sichern und eine Teilhabe an moderner Medizin auch für Menschen in Entwicklungs- und Schwellenländern zu erhalten und zu verbessern. Die globale Teilhabe ist auch eine wichtige Voraussetzung für die Sicherung der Gesundheit in den entwickelten Ländern und für die Entwicklung der medizinischen Praxis und der medizinischen Forschung selbst.

Die Praxis hat gezeigt, dass Open-Source-Technologien die Abhängigkeiten von einzelnen, ausschließlich gewinnorientiert arbeitenden Anbietern verhindert, gleichzeitig Ressourcen spart, weil Neuentwicklungen von Systemen nur einmal und nicht vielfach vollzogen werden müssen und die Qualität bewährter Open-Source-Technologien von den kollektiven Verbesserungen und Erweiterungen einer großen Entwickler-Gemeinschaft profitiert. Hier zwei erfolgreiche Beispiele der jüngeren Vergangenheit:

Beispiel 1: Open-Source-Prinzipien in der Pandemie-Bekämpfung in Taiwan
Die Digitalministerin von Taiwan, Audrey Tang, hat weltweit große Aufmerksamkeit erregt, weil sie mit der Nutzung von Open-Source-Technologien und offener Daten, mit agilen Methoden, Prinzipien des Design Thinkings und der Einbindung der Zivilgesellschaft außerordentliche Erfolge in der Pandemiebekämpfung und Gesundheitsvorsorge in ihrem Land erzielt hat. So wurde z. B. zu Beginn der Pandemie mit Beteiligung der Zivilgesellschaft eine auf Smartphones verfügbare interaktive Karte entwickelt, die den Bürger:innen auf einfachste Weise ermöglichte, Daten über die Verfügbarkeit von Masken in Apotheken und Geschäften abzurufen und selbst einzugeben sowie sich bei landesweiten Verteilmechanismen mit Vorschlägen zu beteiligen. Dies führte zu einer höchst effizienten Versorgung der Bevölkerung mit knappen medizinischen Gütern. Ein anderer sich für die Pandemiebekämpfung als wichtig erweisender Baustein war die schnelle Entwicklung eines sehr einfachen SMS-basierten Check-in-Systems (1922-SMS-Check-in), welches eine wirkungsvolle Kontaktnachverfolgung bei gleichzeitigem Schutz persönlicher Daten ermöglichte.[2]

Die Entwicklung solcher öffentlichen digitalen Services geschieht in Taiwan überwiegend durch gØv („GovZero"), einer zivilgesellschaftlich getragenen Tech Community, an der sich inzwischen mehr als 10.000 Personen, neben Software-Entwick-

2 https://pdis.nat.gov.tw/en/blog/省力安心的簡訊實聯制/

lerinnen und -Entwicklern (40 % der Community) auch öffentliche Bedienstete, Vertreter:innen von Nichtregierungsorganisationen und Bürger:innen aus allen Bereichen der Gesellschaft beteiligen.[3] Audrey Tang kommentierte diese Beteiligung wie folgt: „Das ist wichtig, weil das eine gesellschaftliche Gruppe ist. Sie dient keiner bestimmten Partei und keinem privaten Konzern. Wir dienen den Menschen. Und dabei sind alle Arten von Experimenten erlaubt."[4] Dies führte nicht nur zu wirkungsvollen Instrumenten mit raschen Innovationszyklen, sondern auch zu höchster Akzeptanz der Maßnahmen der Regierung in der Bevölkerung.

Beispiel 2: Corona-Warn-App in Deutschland
Zur Bekämpfung der Corona-Pandemie entstand bereits früh die Idee, über die Bluetooth-Schnittstelle von Mobiltelefonen automatisch zu messen, ob sich die Besitzerin oder der Besitzer eines Telefons über einen zu langen Zeitraum in der Nähe einer infizierten Person aufgehalten hat und somit einem erhöhten Infektionsrisiko ausgesetzt war. Dies ist technisch möglich, aber dazu müssen sehr sensible Informationen möglichst vieler Menschen verarbeitet werden: Historie der Standorte mit Zeitpunkten sowie der Infektionsstatus. Wenn diese Informationen in zentralen Datenbanken gespeichert werden, lassen sie sich nutzen, um jederzeit genaue Bewegungsprotokolle aller Menschen, Informationen zum Zusammentreffen von Personen oder zum Infektionsstatus abzurufen. Es bedarf keiner weiteren Erläuterung, dass solche Daten in staatlicher oder privater Hand mit einem hohen Missbrauchspotential versehen sind.

Offenheit hat technisch und in der Wahrnehmung der App zu einer hohen Sicherheit vor Missbrauch der verarbeiteten Daten und damit zu einer hohen zivilgesellschaftlichen Akzeptanz der App geführt. Experten, wie beispielsweise vom Chaos Computer Club, loben die grundsätzliche Architektur der App und ihre Offenheit regelmäßig; sie ist dadurch in der öffentlichen Wahrnehmung über viele Zweifel erhaben.[5] Dieses in ihrer Offenheit begründete Ansehen der App ist die Voraussetzung für ihre hohe Verbreitung und damit auch für ihren Erfolg. Ohne absolute

3 https://g0v.asia/

4 Audrey Tang im Gespräch mit der Konrad-Adenauer-Stiftung, 07.05.2021: https://www.youtube.com/watch?v=BGGY2tZmPDA

5 Es soll an dieser Stelle dennoch nicht verschwiegen werden, dass es auch berechtigte Kritik an der CWA gibt. Diese betrifft vor allem den Umstand, dass wesentliche Funktionen der App auf gemeinsam von Google und Apple entwickeltem Code beruhen, der Teil des Umfangs der Betriebssysteme Android bzw. iOS ist. Diese Codebestandteile sind zu wesentlichen Teilen nicht offen gelegt. Somit kann letztlich auch für die App als Ganzes nicht ausgeschlossen werden, dass Datenabflüsse insbesondere zu Google oder Apple möglich sind.

Offenheit wäre das Projekt hingegen zum Scheitern verurteilt gewesen, weil eine kritische Masse an Nutzerinnen und Nutzern sie gar nicht installiert hätte und somit Warnungen nicht erfolgt wären. Weil der Programmcode aber auch allen Interessierten in Deutschland zur Verfügung steht, kann er von Forschung, Staat und Wirtschaft auch zur Lösung ähnlicher Probleme in der Zukunft genutzt werden.

Open Source und Open Data in der Medizin

Doch warum sollten die Medizin und das Gesundheitswesen im Allgemeinen auf Offenheit und Open Source setzen? Die UN Sustainable Development Goals unterstreichen bezüglich der Global Health – der Weltgesundheit – die Forderung nach freier Informationstechnologie. Offene Informationssysteme mit offenen Standards und offenen Schnittstellen, die auf einer offen zugänglichen Referenzarchitektur implementiert sind, bieten allen beteiligten Ländern weltweit die Möglichkeit, an diesem wirtschaftlichen Wettbewerb teilzunehmen. Es entstehen gleichwertige Wahlmöglichkeiten für den Nutzer, unabhängig von seinen finanziellen und strukturellen Ressourcen und Rahmenbedingungen. Die Berücksichtigung der genannten Offenheitsaspekte ermöglichen es gerade auch armen Ländern, moderne, passfähige, nachhaltige, robuste Informationstechnik in der Medizin und Forschung mit vergleichsweise geringen Kosten einzusetzen, was wiederum gerade bei der Pandemiebekämpfung auch den entwickelten Ländern zu Gute kommt.

Digitale Souveränität
Persönliche Patientendaten erfordern wegen der Missbrauchsgefahr höchste Ansprüche an den Datenschutz. Trotzdem ist es notwendig und wünschenswert, dass individuelle medizinische Historien einschließlich der Untersuchungsergebnisse und Diagnosen möglichst vollständig und nutzbar zwischen den medizinischen Instanzen ausgetauscht werden können. Davon sind wir noch sehr weit entfernt, und mit fortschreitender Digitalisierung wird diese Aufgabe komplexer, herausfordernder und unübersichtlicher, aber auch chancenreicher. Ein wichtiges Ziel von Open-Source-Entwicklung ist es schon immer gewesen, vertrauenswürdige Software zu entwickeln, die durch Transparenz und Überprüfbarkeit ein Höchstmaß an Sicherheit und Datenschutz gewährleistet und gleichzeitig über offene Standards und Schnittstellen den kontrollierten Datenaustausch ermöglicht. Dass alle Beteiligten dieses Austauschs die volle Kontrolle und den Überblick über ihre Daten bewahren können, ist eine der Kerneigenschaften digitaler Souveränität.

Alle Aspekte wie Offenheit, Transparenz, offener Austausch, Datensouveränität und Datenschutz müssen zusammenspielen, um ihre volle Wirkung zu entfalten. Insbesondere Datensouveränität tritt im Kontext medizinischer Forschung in den Vordergrund und ist eine wichtige Teilmenge digitaler Souveränität. Denn die Möglichkeit für Betroffene zu kontrollieren, wer, wann, zu welchem Zweck etc. auf von einer Person erzeugte oder gespeicherte Daten zugreift und diese wann und wie weiterverarbeiten kann, ist essentiell für die Schaffung von Vertrauen und damit auch für Compliance und valide Daten. Und diese Möglichkeit besteht nur dann, wenn auch die Technologie und die Software kontrolliert werden kann, die diese Daten erzeugt sowie verarbeitet.

Vernetzung
Offenheit und Transparenz sind Voraussetzungen für den wissenschaftlichen Austausch, für Forschung und Innovationen. Globalisierung und Digitalisierung verstärken sich gegenseitig und erhöhen damit den Transformationsdruck in der Medizin. Innovation und Wachstum zeichnen sich durch eine enge Verbindung zwischen Hochschulen, Medizin und Wirtschaft, ein innovations- und investitionsfreundliches Umfeld, allgemein verfügbare und frei nutzbare Technologien, staatliche Investitionen in Zukunftsthemen und offene Menschen aus. Unsere Erfolge und industriellen Stärken der Vergangenheit müssen in der Zukunft auch in den digitalen Kontext übernommen werden. Unsere medizinische Innovationsfähigkeit kann nachhaltig gestärkt werden, wenn Technologie, Software und Daten von Forschenden, Gesundheitsinstitutionen, Unternehmen sowie dem Staat gleichermaßen genutzt werden können. Dabei geht es nicht um Teilungszwang, sondern um die strategische Schaffung von Kooperations- und Mitnutzungsmöglichkeiten. Erfolgreiche Digitalisierung zeichnet sich durch einen hohen Grad an Vernetzung aus. Ziel muss es sein, diese Vernetzung zwischen innovationstreibenden Akteuren weiter zu erhöhen und dabei Software und Daten als Grundlage digitaler Innovationsentwicklung besonders in den Blick zu nehmen. Dies bildet die Basis dafür, dass Forschung parallel von verschiedenen Seiten und Akteuren vorangetrieben wird und damit viel schneller Effizienzpotentiale gehoben werden können.

Unabhängigkeit
Die Erfahrung hat gezeigt, dass technologische Abhängigkeit von einzelnen Anbietern große Gefahren birgt. Monopolstellungen gefährden ganz besonders im digitalen Bereich die Souveränität von Menschen, Unternehmen, Institutionen und Nationen. Gerade in der Medizin ist es sehr wichtig, solche Abhängigkeiten zu verhindern. Denn digitale Souveränität bedeutet auch, sich nicht in zu starke Abhängigkeiten einzelner Akteure zu begeben. Egal, ob es sich um Software,

IT-Hardware, medizinisches Equipment oder medizinisches Forschungsinfrastruktur handelt: Herstellung, Lieferung und Prüfung müssen immer auch durch unabhängige Instanzen ohne kommerzielle Interessen möglich sein und zukünftig auch regelmäßig erfolgen. Transparenz ist hierbei das zentrale Kriterium. Transparenz in der Technik zeigt sich darin, wie sie gebaut oder wie Software programmiert ist, wie sie Daten erzeugt, verarbeitet und auswertet. Und das gilt unabhängig davon, ob es sich um Daten für Informationsbedarfe handelt oder um Daten, die über Menschenleben entscheiden, weil sie beispielsweise die medizinische Forschung oder Entscheidungen für Medizinzulassungen beeinflussen.

Sicherheit der Infrastruktur
Im Juli 2019 mussten sich DRK-Krankenhäuser in Rheinland-Pfalz und im Saarland gegen einen Hackerangriff zur Wehr setzen. Dabei war das komplette Netzwerk des Verbands Süd-West betroffen. Zu Beginn 2021 wurde die Urologische Klinik in Planegg im Süden Münchens Ziel eines Cyberangriffs und im März die Evangelische Klinik in Lippstadt. Die Liste wird immer länger, die Sicherheitssysteme sind fragiler, als viele hoffen. Ein Ausfall der IT oder vernetzter Medizintechnik kann lebensbedrohlich sein – dies darf bei einem Teil der kritischen Infrastruktur einfach nicht passieren. Trauriges und zugleich mahnendes Beispiel ist die Cyberattacke auf die Uniklinik Düsseldorf, die eine lebensbedrohlich erkrankte Patientin abweisen musste, weil die Systeme nicht funktionierten. Die Frau musste in ein anderes Krankenhaus verlegt werden und die Behandlung konnte erst eine Stunde später und somit zu spät beginnen. Dies hatte zur Folge, dass die Patientin verstarb.

Dass gerade bei kritischen Infrastrukturen die Transparenz und Dynamik von Open-Source-Technologien zu extrem schnellen Lösungen führen kann, hat der Sicherheitsvorfall um Log4J gezeigt. Log4J ist eine weit verbreitete Protokollierungsbibliothek, bei der im Dezember 2021 eine sicherheitsrelevante Schwachstelle entdeckt wurde. Über diese Schwachstelle hätten Hacker relativ einfach in Systeme eindringen können. Weil es sich bei Log4J um eine Open-Source-Software handelt, ist eine große Community an der Entwicklung beteiligt. Dieser Community ist es in atemberaubender Geschwindigkeit gelungen, die Schwachstelle zu entfernen und eine überarbeitete Version zur Verfügung zu stellen. Natürlich kann auch Open-Source-Software-Schwachstellen enthalten, aber die Gefahr ist bei weit verbreiteten Open-Source-Werkzeugen geringer, weil sehr viele Beteiligte den Code nicht nur sehr genau kennen, sondern ständig prüfen und eventuelle Fehler sehr viel schneller beheben können, als das in Unternehmen mit einem vergleichsweise kleinen Entwicklerteam möglich wäre.

Medizinische Innovation souverän gestalten

Um eine zukunftsfähige Digitalisierung in der Medizin zu ermöglichen, empfehlen wir ein konsequentes Zusammenspiel auf verschiedenen Ebenen von Offenheit:

Open-Source-Software
Software für die medizinische Forschung sollte stets unter einer Open-Source-Lizenz entwickelt und veröffentlicht werden. Wenn medizinische Forschung zur Erreichung von Ergebnissen gezielt Software erstellt und verwendet, muss gewährleistet werden, dass diese Software als Teil des Forschungsprozesses und seiner Ergebnisse selbst betrachtet wird und damit dieser Maxime genügen muss. Hierzu ist es unabdingbar, dass Einblick in den Quellcode von Software genommen werden kann, um ihn hinsichtlich der Methodik und Integrität der Verarbeitung zu überprüfen und einer wissenschaftlich kritischen Würdigung zuführen zu können. Diese Forderung wird in dem Maße wichtiger, in dem die Software selbst wesentliche neue Erkenntnisse gewinnt, anstatt lediglich mittelbar hilfreich zu sein. Wenn also Software erstellt wird, zum Beispiel um gezielt neuronale Netze zu trainieren, die wiederum selbst zur Generierung von neuen Informationen und Zusammenhängen verwendet werden, so unterliegt diese dem Gebot nach Open Source noch sehr viel stärker als Software, die lediglich dazu dient, Ergebnisse festzuhalten und zu präsentieren.

Hinzu kommt: Wenn die Ergebnisse öffentlich finanzierter Forschung der Allgemeinheit zur Verfügung stehen sollen, wirkt die wirtschaftliche Förderung von Open Source als Innovationstreiber und gewährleistet, dass die Investitionen auch wieder der Gesellschaft zugutekommen. Gleichzeitig vereinfacht Open Source die Weitergabe und Nutzung von Software und Daten. Durch die Wahl geeigneter Lizenzmodelle kann der öffentliche und gemeinnützige Sektor sogar den Gewinn an Erkenntnissen und Forschungstätigkeiten vergrößern, da Dritte an der Weiterentwicklung ohne den Verlust von Zeit und Ressourcen teilnehmen und die Forschung nahtlos weiterführen können, indem sie direkt auf Bestehendem aufbauen. Das Beispiel des Betriebssystems Linux als Vertreter eines Open-Source-Entwicklungsmodells hat gezeigt, wie eine Open Source basierte Softwareentwicklung eine Dynamik und Breite erzeugt, der andere Modelle wie zum Beispiel Windows oder MacOS nichts entgegensetzen können. Diese Dynamik und Breite kommt der Forschung direkt zugute und erzeugt einen greifbaren Mehrwert für die Gesellschaft, indem es den wissenschaftlichen Erkenntnisgewinn von Forschung beschleunigt und verbreitert.

Open Data

Das E-Government-Gesetz vom Juli 2017[6] hat die Weichen für eine umfangreiche Öffnung der Verwaltungen und deren Daten gestellt. Unter anderem begünstigen §12 und insbesondere §12a die Veröffentlichung von Daten. Durch die öffentliche Verfügbarkeit der Daten soll die Teilhabe an Informationen aus Regierung und Verwaltung befördert werden und neben der einfachen Offenlegung auch die Schaffung neuer davon abgeleiteter Projekte gefördert werden. Unter dieses Veröffentlichungsgebot fallen im Grunde alle durch die öffentliche Hand finanzierten Datensammlungen und entsprechend auch durch die öffentliche Hand finanzierte Forschungstätigkeit. Diese Vorteile treffen im Grunde auch auf durch medizinische Forschung erzeugte Daten zu. Daten können einer breiteren Nutzung zugeführt werden und die Verknüpfung von Daten über einzelne Projekte hinweg ermöglicht zusätzliche Erkenntnisse. Open Source ist in besonderem Maße dafür geeignet, Open Data anzubieten, da die Offenheit der Verarbeitung und Nutzung von Daten im Kern bereits selbst angelegt ist.

Die Open Source Business Alliance hat sich 2020 beim letzten Digitalgipfel der Bundesregierung gemeinsam mit vielen anderen mit den Themen Nachhaltigkeit, Resilienz und digitale Souveränität intensiv in die Fokusgruppe Digitale Souveränität eingebracht. Auch hier wurden explizite Forderungen und Handlungsfelder für die Bildung und Wissenschaft abgeleitet. Software und Software-Plattformen für die medizinische Forschung sollten auf Open Standards, Open Platforms und Open Source bauen, um die Freiheit und Wissenschaftlichkeit der Forschung sicherzustellen und davon abgeleitet eine breitere Forschungstätigkeit und eine höhere Dynamik zu fördern.

Open Standards

Die Forderung nach offenen Standards (Open Standards) ist keine, die sich ausschließlich aus dem Umfeld von Software ergibt, sondern sie begegnet uns schon sehr früh auf dem Wege wirtschaftlicher und sozialer Entwicklung. In vielen Städten in Europa finden sich vor allem in der Nähe von Marktplätzen oftmals in das Mauerwerk eingelassene Maße. So konnten Käufer und Verkäufer sich klar an einem offenen Standard orientieren, etwa um prüfen zu können, wie groß konkret eine Elle sein sollte und ihre Erwartungen und Interaktionen darauf hin abstimmen. Derartige offene Standards befrieden das Verhältnis der Akteure und schaffen Sicherheit und Vertrauen. Übertragen auf Software bedeutet das, dass

6 https://www.bundestag.de/resource/blob/655082/32a17c3834d5c5c5d6f5a7232f0491c0/WD-3-134-19-pdf-data.pdf, 2019.

Datenquellen ohne Verfälschung oder unlautere Abschirmung interaktiv und wechselseitig verarbeitet werden können.

Insbesondere spezielle Datenformate können ohne einen direkten oder erkennbaren Gegenwert eine nachhaltige und tiefgreifende Bindung an eine explizite Software bzw. einen bestimmten Software-Hersteller erzeugen. Wenn zum Beispiel bereits eine umfangreiche Datensammlung vorliegt, diese aber nur mit einer bestimmten Software lesbar ist, dann muss unter Umständen an dieser Software festgehalten werden, obwohl andere Anbieter fortgeschrittenere oder günstigere Software anbieten. Offene Standards (Open Standards) vermeiden diese Problematik und Open Source ist gegen eine solche Entwicklung immun, da der Quellcode des Datenformats offenliegt und entsprechend genutzt werden kann. Der sogenannte „Vendor-lock-in" wird effektiv verhindert.

Offene Plattformen (Open Platforms)
Was für offene Standards gilt, gilt in gleichem Maße auch für Plattformen. Unter Software-Plattformen versteht man im Allgemeinen eine einzelne oder eine Sammlung von Software und Dienstleistungen, die von einer kleinen Gruppe oder einem einzelnen Unternehmen kontrolliert werden. Meist starten diese Plattformen mit zunächst kostenlosen Angeboten, mit zunehmender Bedeutung und steigender Marktmacht werden diese jedoch zunehmend restriktiver geführt und die weitere Entwicklung stärker gesteuert und kontrolliert. Häufig wird dabei das Ziel verfolgt, Daten, die durch die Plattform-Teilnehmer erzeugt oder anderweitig erbracht werden, ohne deren Kenntnis abzuschöpfen und in eine wirtschaftliche Nutzung zu überführen. Dieses Gebaren bedroht direkt die Freiheit der Forschung und somit auch der medizinischen Forschung und bietet potentiell die Möglichkeit, der Forschung selbst ihren erzeugten Mehrwert vorzuenthalten.

Eine solche Eigendynamik kann durch offene Plattformen vermieden werden. Im besten Fall stellen offene Plattformen ein Ökosystem bereit, in dem die Teilnehmer einfach und unkompliziert einzelne Dienstleistungen für andere Teilnehmer oder sonstige Nutzer erbringen können, die wiederum das Angebot der offenen Plattform erweitern und verbessern. Ein Beispiel ist das Gaia-X-Projekt, das sich anschickt, eine solche offene Plattform zu beschreiben und zu motivieren. Darüber hinaus ist die Nationale Forschungsdateninfrastruktur (NFDI) ein Beispiel für einen offenen Wissensspeicher.[7]

7 Vgl. den Artikel zur NFDI von Iris Pigeot et al. in diesem Heft.

Gaia-X

Mit der Initiative Gaia-X wird ein offenes, transparentes und sicheres europäisches Daten- und Infrastruktur-Ökosystem geschaffen, das höchsten Anforderungen an digitaler Souveränität genügt und in dem Daten sicher und vertrauensvoll verfügbar gemacht, zusammengeführt und geteilt werden können. Gaia-X soll dazu beitragen, Europa unabhängiger zu machen und die technologische Souveränität im Umgang mit den Daten von Bürger:innen und privaten wie öffentlichen Institutionen zu stärken. Institutionen, Unternehmen und Bürgerinnen und Bürger sollen Daten sammeln und weitergeben können – und zwar so, dass sie die Kontrolle darüber behalten. Sie sollen entscheiden können, was mit ihren Daten geschieht, wo sie gespeichert werden, und dabei stets die Datenhoheit behalten.

Die Architektur von Gaia-X basiert auf den Prinzipien von Dezentralisierung, Offenheit, Transparenz und Vertrauen durch gemeinsam definierte Standards, den Gaia-X-Standards. Gaia-X ist in nationalen Hubs organisiert. Zu den Projekten der Working Group Health des German Gaia-X-Hub zählen: AIQNET (Data collection and AI methods for the automated extraction and analysis of data), Berlin Health Data Space (AI to beat acute kidney failure), Smart Health Connect (Preventative healthcare with smart wearables), Research Platform Genomics (Research cloud for genome data to defeat cancer) u. v. a. m.[8]

Nationale Forschungsdateninfrastruktur (NFDI)

Ein dauerhafter, allgemein zugänglicher, hochqualitativer digitaler Wissensspeicher und darauf basierende Dienste werden als unverzichtbare Voraussetzung für neue Forschungsfragen, Erkenntnisse und Innovationen angesehen. Mit der „Nationalen Forschungsdateninfrastruktur" (NFDI)[9], die im Oktober 2020 als Institution gegründet wurde, bauen Bund und Länder eine vernetzte, interoperable, nachhaltige Informationsinfrastruktur auf. Wertvolle Datenbestände von Wissenschaft und Forschung sollen für das gesamte deutsche Wissenschaftssystem systematisch erschlossen, vernetzt und nachhaltig nutzbar gemacht werden. Die NFDI ist auch Teil internationaler Initiativen wie z. B. der European Open Science Cloud (EOSC)[10]. Die Dateninfrastrukturen folgen den sog. FAIR-Prinizipien: Findable,

8 https://www.data-infrastructure.eu/Redaktion/EN/Dossier/gaia-x.html#doc2845524bodyText7

9 https://www.nfdi.de/

10 https://eosc-portal.eu/

Accesible, Interoperable, Reusable. Der Aufbau der Dateninfrastrukturen wird in Konsortien mit öffentlichen Mitteln von derzeit bis zu 90 Mio. € pro Jahr gefördert. Zu den Medizin-relevanten Konsortien zählen u. a. NFDI4Health (personenbezogene Gesundheitsdaten) und GHGA (Deutsches Humangenom-Phänom-Archiv).

Mit der Initiative FAIR-Data Spaces[11] soll ein gemeinsamer Cloud-basierter Datenraum für Wissenschaft und Wirtschaft „so frei und offen wie möglich“[12] geschaffen werden durch die Verknüpfung von NFDI mit Gaia-X. Langfristiges Ziel ist es, die nationale Architektur zu einer europaweiten und globalen Infrastruktur zu erweitern, basierend auf europäischen Normen und Werten, um digitale Souveränität sicherzustellen.

EU-Studie zu Open Source

Eine von der EU-Kommission in Auftrag gegebene umfassende Studie[13] belegt einen signifikanten Einfluss von Open Source auf die Wettbewerbsfähigkeit europäischer Unternehmen, auf das Wirtschaftswachstum, auf die Start-up-/KMU-Szene und die technologische Unabhängigkeit. Die EU-Kommission empfiehlt daher ihren Mitgliedsstaaten, Open Source auf allen Ebenen von der Bildung über die Forschung, den öffentlichen Sektor bis hin zur Wirtschaftspolitik zu fördern. Bezüglich der Forschung wird empfohlen, die Entwicklung von Open-Source-Software und -Hardware und die Förderung von Open-Source-Communities in die Forschungs- und Innovationspolitik, auch in bestehende Programme, wie z. B. Horizon Europe, konsequent zu integrieren und durch entsprechende Mittel zu fördern. Open Source sollte als wesentlicher Bestandteil von Wissens- und Technologietransfer verstanden werden, z. B. als explizite Transferanforderung für öffentlich geförderte Programme. Es sollen wirksame Anreize für die Veröffentlichung von in Forschungs- und Entwicklungsprojekten entstandenem Code in öffentlich zugänglichen EU-basierten Open-Source-Repositorien geschaffen werden. Hochschulen und öffentlich finanzierte Forschungseinrichtungen sollen unternehmerische Fähigkeiten und Open-Source-Kenntnisse in allen relevanten Curricula vermitteln und spezielle Aufbaustudiengänge anbieten.

11 https://www.nfdi.de/fair-data-spaces/

12 Prof. Dr. York Sure-Vetter (https://www.nfdi.de/fair-data-spaces/, 01.01.2022).

13 https://digital-strategy.ec.europa.eu/en/library/study-about-impact-open-source-software-and-hardware-technological-independence-competitiveness-and
Eine deutschsprachige Zusammenfassung finden Sie hier: https://osb-alliance.de/verbands-news/eu-studie-open-source-staerkt-die-wirtschaft-und-die-technologische-unabhaengigkeit

Fazit

Offenheit entspricht der wissenschaftlichen Vorgehensweise, ohne Offenheit hätten Forschung und Lehre niemals das heutige, hohe Niveau erreichen können. Die Entwicklung des Internets zeigt, wie sehr offene Systeme, um die sich die Wissenschaft seit jeher bemüht und an denen sie maßgeblich beteiligt war, den weltweiten Fortschritt durch die Verbreitung von Wissen vorangebracht haben. Die Digitalisierung selbst ist eine Folge dieses Fortschritts, doch sie stellt uns vor große Herausforderungen. Wissenschaftliche Forschung und insbesondere die Medizin erfordern höchste Standards in Bezug auf offenen Wissenstransfer, sichere Infrastrukturen, Datenschutz und digitale Souveränität. In der Medizin behindern proprietäre Technologien die Entwicklung von Möglichkeiten. Das führt zu massiven Mehrfachaufwänden mit jeweils dem gleichen Ergebnis, benachteiligt insbesondere wirtschaftlich schwächere Länder und stellt damit eine massive Gefahr für die Weltgesundheit dar. Die Corona-Pandemie hat uns gezeigt, dass Gesundheit immer weltweit betrachtet werden muss und alle Länder in die Lage versetzt werden müssen, die Gesundheit der Menschen zu schützen. Sonst können sich Pandemien entwickeln, die auch in den entwickelten Ländern nicht kontrollierbar sind.

Open-Source-Technologien sind im Umfeld der Wissenschaft entstanden, weil der sichere Austausch von Information einen kollektiven Wert darstellt, der nicht durch Einzelinteressen behindert, blockiert oder gar missbraucht werden darf. Über Jahrzehnte hat die Open-Source-Szene Konzepte entwickelt und umgesetzt, die heute zum einen bewährte Standards sind und zum anderen den größten Bedrohungen entgegenstehen, die die fortschreitende Digitalisierung für die individuelle und kollektive Freiheit darstellt. Open Source bedeutet gemeinsame Entwicklung, freien Austausch und ein Höchstmaß an Sicherheit. Als Open Source Business Alliance bringen wir uns gerne in den Prozess ein, Open Source in der Medizin zum Standard zu etablieren.

VERANTWORTUNGSVOLLE SEKUNDÄRNUTZUNG VON PATIENTENDATEN

Daniel Strech

Einleitung

Das primäre Interesse jeder Form von Forschung sollte der Erkenntnisgewinn sein.[1] Ein entsprechender Erkenntnisgewinn im Kontext biomedizinischer Forschung kann sich auf ein besseres Krankheitsverständnis, auf Ansätze zur Therapie von Krankheiten und viele weitere Bereiche beziehen. Neben präklinischer Forschung und klinischen Studien, birgt auch die Sekundärnutzung von Patientendaten ein hohes Potential für einen biomedizinischen Erkenntnisgewinn.

Allerdings stellen nicht alle Ergebnisse von Forschungsprojekten immer einen Erkenntnisgewinn dar. Forschungsprojekte können bewusst oder unbewusst so beeinflusst werden, dass ihre Ergebnisse ein verzerrtes Abbild der Wahrheit wiedergeben.[2] Dann handelt es sich nicht um einen Gewinn, sondern eher um eine Verzerrung von Erkenntnissen, denen man nicht vertrauen kann. Aber auch bei nicht verzerrten Ergebnissen kann der Erkenntnisgewinn stark eingeschränkt sein, wenn er für die Wissenschaft nicht zugänglich ist, weil er zum Beispiel sprichwörtlich in der Schublade verschwindet.[3] Die Nützlichkeit von Erkenntnissen steigt mit deren Zugänglichkeit bzw. Offenheit (Stichwort: Open Science). Neben der Offenheit erhöht sich der Nutzen von Forschung weiter, wenn sie besonders Patient:innen-orientierte Fragen adressiert.[4] Nicht zuletzt müssen Forschungsprojekte ethisch vertretbar sein, damit diese Form des Erkenntnisgewinns von der Gesellschaft akzeptiert und finanziert wird.

1 Vgl. BVerfGE 35, 79 – Niedersächsisches Hochschulgesetz.

2 Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. 2014. Increasing value and reducing waste in research design, conduct, and analysis. Lancet, 383(9912): 166–175.

3 Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gotzsche PC, Krumholz HM, Ghersi D, van der Worp HB. 2014. Increasing value and reducing waste: addressing inaccessible research. Lancet, 383(9913): 257–266.

4 Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gulmezoglu AM, Howells DW, Ioannidis JP, Oliver S. 2014. How to increase value and reduce waste when research priorities are set. Lancet, 383(9912): 156–165.

Ob eine Wissenschaft und somit auch die Sekundärnutzung von Patientendaten verantwortungsvoll einen Erkenntnisgewinn generiert, lässt sich also daran festmachen, wie vertrauenswürdig, nützlich (für die Wissenschaft und Patient:innen) und ethisch sie ist.[5] Soweit die Theorie, aber wie lässt sich konkretisieren und evaluieren, ob beispielsweise die Sekundärnutzung von Patientendaten Erkenntnisse in mehr oder weniger vertrauenswürdiger, nützlicher und ethischer Form generiert?

Prozeduren zur Förderung von Vertrauenswürdigkeit, Nützlichkeit und Ethik

Für klinische Studien und zunehmend auch für Tierstudien wurden in den letzten Jahrzehnten verschiedene Prozeduren definiert, welche die Vertrauenswürdigkeit, Nützlichkeit und Ethik individueller Forschungsprojekte konkret verbessern. So lässt sich die Vertrauenswürdigkeit von Studien zum Beispiel durch Fallzahlberechnung, Randomisierung, Verblindung der Endpunkterhebung oder durch die Replikation von Studien erhöhen.[6] Auf diese Weise lassen sich die Transparenz von Studien und die Zugänglichkeit ihrer Ergebnisse fördern. Zu diesen Prozeduren zählen die Registrierung von Studienvorhaben vor dem Studienstart, eine zeitnahe Ergebnisveröffentlichung oder Data Sharing. Die Nützlichkeit des Erkenntnisgewinns im Sinne seiner praktischen Relevanz kann beispielsweise durch die von Patient:innen und behandelnden Ärzt:innen gemeinsam entwickelten patienten-relevanten Fragestellungen verbessert werden. Die ethische Seite kann ergänzend durch solche Prozeduren verbessert werden, die den Schutz von Studienteilnehmenden fördern. Hierzu gehören u. a. unabhängige Nutzen-Schaden-Abwägungen durch Ethikkommissionen, Datenschutz und die informierte Einwilligung von Studienteilnehmenden.[7]

5 Strech D, Weissgerber T, Dirnagl U, Group Q. 2020. Improving the trustworthiness, usefulness, and ethics of biomedical research through an innovative and comprehensive institutional initiative. PLoS Biol, 18(2): e3000576.

6 Schmucker C, Nothacker M, Rücker G, C. M-B, Kopp I, Meerpohl J. 2016. Bewertung des Biasrisikos (Risiko systematischer Fehler) in klinischen Studien, https://www.cochrane.de/sites/cochrane.de/files/uploads/manual_biasbewertung.pdf

7 World Medical Association. 2013. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects, Fortaleza.

Best-Practice-Standards

Die Festlegung entsprechender Prozeduren ist jedoch nur der erste Schritt. Individuelle Interpretationen davon, was eine angemessene Randomisierung, Registrierung oder Patientenbeteiligung ist, können sich stark unterscheiden. Für jede dieser oben genannten Prozeduren haben sich deshalb im Kontext klinischer Studien und zunehmend auch für Tierstudien Best-Practice-Standards etabliert, die dynamisch weiterentwickelt werden. Zum Beispiel wurde spezifiziert, wann eine Randomisierung mehr oder weniger relevant ist[8] und wie diese mit hoher Qualität durchgeführt werden kann.[9] Die WHO hat spezifiziert, was unter „zeitnaher" Ergebnisveröffentlichung konkret zu verstehen ist.[10] Diese Best-Practice-Standards informieren und unterstützen nicht nur die Forschenden selber, sondern erlauben den akademischen Institutionen, Forschungsförderern, Fachzeitschriften und weiteren Akteur:innen eine valide empirische Evaluation dahingehend, welche Prozeduren verantwortungsvoller Wissenschaft mehr oder weniger gut umgesetzt werden: So wurde versucht, die Frage zu beantworten, wie oft in präklinischen Studien randomisiert wird,[11] und auch wie oft und wie zeitnah die Ergebnisse abgeschlossener klinischer Studien veröffentlicht werden.[12] Die Ergebnisse dieser empirischen Evaluationen oder Meta-Research-Studien erlauben ein Monitoring und eine Qualitätssicherung dieser wichtigen Forschungsfelder. Darüber hinaus zeigen entsprechende Evaluationen auch auf, ob Verzerrungen in den veröffentlichen Studienergebnissen vorliegen z. B. dadurch, dass positive/erwünschte Ergebnisse häufiger veröffentlicht werden als negative/unerwünschte.

8 Kimmelman J, Mogil JS, Dirnagl U. 2014. Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol, 12(5): e1001863.

9 Rosenberger WF, Lachin JM: Randomization in clinical trials. 2016. Theory and practice. Hoboken, New Jersey: John Wiley & Sons, Inc.

10 WHO. 2017. Joint statement on public disclosure of results from clinical trials; www.who.int/ictrp/results/jointstatement/en/

11 Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, Hirst T, Hemblade R, Bahor Z, Nunes-Fonseca C et al. 2015. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol, 13(10): e1002273.

12 Riedel N, Wieschowski S, Bruckner T, Holst MR, Kahrass H, Nury E, Meerpohl JJ, Salholz-Hillel M, Strech D. 2021. Results dissemination from completed clinical trials conducted at German university medical centers remained delayed and incomplete. The 2014 -2017 cohort. J Clin Epidemiol, 144: 17.

Zur Sekundärnutzung von Patientendaten

Für die daten-getriebene Forschung, die auf der Sekundärnutzung von Patientendaten beruht, hat die grundlegende Klärung der relevanten Prozeduren und entsprechend die Etablierung von Best-Practice-Standards für vertrauenswürdige, nützliche und ethische Wissenschaft gerade erst begonnen. So müssen noch die folgenden Fragen geklärt werden: Wie lässt sich die Vertrauenswürdigkeit/Robustheit von Ergebnissen aus der Sekundärnutzung konkret operationalisieren? Welche Anforderungen sollten an die Reproduzierbarkeit von Sekundärnutzungen gestellt werden? Welche Transparenz bei Sekundärnutzungen verbessert deren Nützlichkeit? Welche Form von informierter Einwilligung ist zu bevorzugen? Da selbst diese zum Teil sehr fundamentalen Prozeduren in vielen Bereichen der Sekundärnutzung von Patientendaten noch nicht geklärt bzw. nicht breit konsentiert sind, gibt es auch kaum empirische Evaluationen (Meta-Research) zum Status quo. Entsprechend lassen sich auch nur begrenzt objektive Aussagen darüber treffen, inwieweit die veröffentlichten Ergebnisse aus der Sekundärnutzung zu Verzerrungen neigen oder nicht.

Dieser Beitrag skizziert einige bislang nicht ausreichend geklärte Prozeduren für eine verantwortungsvolle Sekundärnutzung. Für einen Überblick siehe Tabelle 1. Der Fokus liegt dabei auf den beiden Bereichen Vertrauenswürdigkeit und Nützlichkeit. Die Fragen zur Ethik der Sekundärnutzung und hier insbesondere die Fragen nach dem geeignetsten Modell für die Einwilligung zur Datenspende (broad consent, dynamic consent, opt-out) wurden bereits an anderen Stellen intensiv diskutiert.[13] Des Weiteren werden empirische Fragen der Qualitätssicherung/Meta-Research skizziert, die bearbeitet werden können, wenn die Prozeduren und damit zusammenhängende Best-Practice-Standards grundsätzlich geklärt sind. Diese Überlegungen erheben keinen Anspruch auf Vollständigkeit. Ziel dieses Textes ist die Erhöhung des Problembewusstseins für diesen Aufgabenbereich, der insgesamt eine zentrale Komponente der Governance lokaler, nationaler und internationaler Sekundärnutzung von Patientendaten ist.

13 Siehe z. B. zur deutschen Diskussion die Unterlagen der Medizininformatik-Initiative (MII) www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung oder das Wissenschaftliche Gutachten „Datenspende", www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Ministerium/Berichte/Gutachten_Datenspende.pdf

Vertrauenswürdigkeit

Die Vertrauenswürdigkeit von Ergebnissen aus der Sekundärnutzung hängt unter anderem ab von a) der Qualität der zugrundeliegenden Patientendaten[14], b) den Maßnahmen zur Reduktion von Verzerrungen (Bias) in den Datenanalysen[15] oder auch c) der Reproduzierbarkeit der Ergebnisse.[16] Für alle diese drei Bereiche bedarf es der Operationalisierung von Prozeduren und Best-Practice-Standards: Wie lässt sich die Vertrauenswürdigkeit bestimmter Patientendaten und der angewendeten Analysetechniken als niedrig, mittel oder hoch operationalisieren und wie sähe ein konsensfähiger Best-Practice-Standard diesbezüglich aus? Was wären Mindestanforderungen an die Reproduzierbarkeit der Ergebnisse aus Sekundärnutzungen?

Ähnlich wie in der klinischen Forschung wird man auch im Bereich der Sekundärnutzung solche Prozeduren auf bestimmte Studien und deren Fragestellungen hin spezifizieren müssen. Zugleich muss es aber eine Ebene von Best-Practice Standards geben, die nicht zu kleinteilig ist, da sonst eine für die verschiedenen Stakeholdergruppen verständliche Qualitätsbewertung und übergreifende Evaluationen zum Status quo kaum möglich sind. Nicht nur die Expert:innen für Data Science und Medizininformatik sollten ein konsensfähiges Verständnis für ihre oft sehr speziellen Fragestellungen haben, sondern entsprechende Standards müssen auch für weitere Stakeholder wie Ärzt:innen, Krankenversicherungen, Regulierungsinstanzen und nicht zuletzt Patient:innen per se verständlich bleiben, wobei eine gewisse Einarbeitung in die Materie vorausgesetzt wird. Für klinische Studien gibt es eine solche Situation u. a. durch die Verfügbarkeit von Reporting Guidelines wie CONSORT.[17] Heutzutage ist vielen Ärzt:innen und auch informierten Patient:innen bewusst, dass eine klinische Studie durch Fragen zur Fallzahl, Studienpopulation, Randomisierung etc. grob auf ihre Vertrauenswürdigkeit hin bewertet werden kann. Eine solche orientierende Bewertung der Vertrauenswürdigkeit durch „relative Laien" ist wichtig, weil diese Personengruppen am Ende Entscheidungen über die Anwendung (oder Finanzierung) der in den Studien untersuchten Diagnose-, Präventions- oder Therapie-Ansätze treffen müssen. Für Ergebnisse der Sekundär-

14 Zulman DM, Shah NH, Verghese A. 2016. Evolutionary Pressures on the Electronic Health Record: Caring for Complexity. JAMA, 316(9): 923–924.

15 Lazer D, Kennedy R, King G, Vespignani A. 2014. Big data. The parable of Google Flu: traps in big data analysis. Science, 343(6176): 1203–1205.

16 Beam AL, Manrai AK, Ghassemi M. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA, 323(4): 305–306.

17 Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med, 7(3):e1000251.

nutzung ist ein solches Grundverständnis bislang nur schwer möglich, weil entsprechende Best-Practice Standards fehlen.

Nützlichkeit

Auch bei der Sekundärnutzung von Patientendaten steigt die Nützlichkeit des Erkenntnisgewinns, wenn die Ergebnisse vollständig, zugänglich und verständlich veröffentlicht werden. Im Kontext klinischer Studien empfehlen internationale Leitlinien wie die Deklaration von Helsinki, dass alle klinische Studien vor ihrem Start öffentlich zugänglich registriert und dass alle Ergebnisse (auch negative) veröffentlicht werden sollen.[18] Ergänzende Empfehlungen der WHO spezifizieren, dass diese Veröffentlichung innerhalb von zwei Jahren nach Studienende geschehen sollte und dass zusätzlich Kurzzusammenfassungen auf der Registerseite zur Verfügung gestellt werden sollen.[19] Verschiedene Meta-Research-Studien konnten auf der Basis dieser Best-Practice-Standards untersuchen, wo die Nützlichkeit klinischer Studien ein hohes Level erreicht hat und wo sie noch stärker ausgebaut werden müsste.[20]

Für die Sekundärnutzung von Patientendaten fehlen gegenwärtig diese Standards und somit lassen sich aktuell auch nur sehr beschränkte Aussagen zu Stärken und Schwächen ihrer Nützlichkeit machen. In Analogie zur Deklaration von Helsinki für klinische Studien hat der Weltärztebund z. B. die Deklaration von Taipeh zur Forschung mit Gesundheitsdaten und Biobanken veröffentlicht.[21] Diese betont zwar die Wichtigkeit von Transparenz, beschreibt aber anders als die Deklaration von Helsinki keine spezifischen Prozeduren zu deren Umsetzung. Die Deklaration

18 World Medical Association. 2013. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects, Fortaleza; Riedel N, Wieschowski S, Bruckner T, Holst MR, Kahrass H, Nury E, Meerpohl JJ, Salholz-Hillel M, Strech D. 2021. Results dissemination from completed clinical trials conducted at German university medical centers remained delayed and incomplete. The 2014–2017 cohort. J Clin Epidemiol, 144: 17.

19 WHO. 2017. Joint statement on public disclosure of results from clinical trials; www.who.int/ictrp/results/jointstatement/en/

20 Chan AW et al. 2014. Increasing value and reducing waste: addressing inaccessible research. Lancet, 383(9913):257-266; Riedel N et al. 2021. Results dissemination from completed clinical trials conducted at German university medical centers remained delayed and incomplete. The 2014 – 2017 cohort. J Clin Epidemiol, 144: 17.

21 World Medical Association (WMA). 2016. Declaration of Taipei on ethical considerations regarding health databases and biobanks In. https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/

von Taipeh wie auch andere Beiträge zum Thema Sekundärnutzung nennen meist nur die Festlegungen von Komitees (Use & Access Committees) und Vereinbarungen (Use & Access Policies) für den Zugriff auf Daten. Konkrete Prozeduren wie die Registrierung und Ergebnisveröffentlichung als mögliche Best-Practice-Standards werden nicht erwähnt. Bei einigen wenigen Datenbanken für die Sekundärnutzung findet man allerdings bereits entsprechende Spezifizierungen. So legen z. B. die neu geschaffene Infrastruktur und die Zugangsregeln für die Sekundärnutzung der vom Institut für Qualität und Transparenz im Gesundheitswesen (IQTiG) zur Qualitätssicherung erhobenen Daten fest, dass alle Anträge zur Sekundärnutzung öffentlich einsehbar sein sollen und eine Veröffentlichung der Ergebnisse innerhalb von zwei Jahren nach Erhalt der Daten erwartet wird.[22]

Im Bereich der klinischen Forschung wurden auch Best-Practice-Standards zur Förderung der praktischen Relevanz des Erkenntnisgewinns entwickelt. Dazu zählen Prozeduren, um Patient:innen beim Design klinischer Studien und bei der Auswahl relevanter Fragestellungen zu beteiligen.[23] Auch bei der Festlegung von Patienten-relevanten Zielgrößen (sogenannte Core Outcome Sets (COS)) werden entsprechende Prozeduren beschrieben.[24] Entsprechend wäre auch für die Sekundärnutzung von Patientendaten zu klären, welche Prozeduren von Patienten-/Bürgerbeteiligung die praktische Relevanz der Sekundärnutzung effektiv fördern können. Entsprechende Prozeduren wären sowohl bei der Begutachtung individueller Projekte vorstellbar als auch bei einer eher übergreifenden Evaluation zur Performance der Sekundärnutzung.

In diesem Zusammenhang gewinnt auch eine ergänzende Form der Transparenz zur Sekundärforschung an Bedeutung. Es wäre bei durchgeführten Projekten (Sekundärnutzungen) und deren Ergebnissen zu überlegen, ob die durch Use & Access-Komitees abgelehnten Anfragen zur Sekundärnutzung veröffentlicht werden sollten. Die dadurch erzeugte Transparenz würde es ermöglichen, die übergreifende Nützlichkeit der Sekundärnutzung von Patientendaten zu evaluieren. Es könnte z. B. sein, dass insbesondere solche Anfragen negativ beschieden werden, die konkrete Fragen der Versorgungsforschung oder Qualitätssicherung untersuchen, weil Krankenhäuser, in denen die jeweiligen Patient:innen behan-

22 https://iqtig.org/datenerfassung/sekundaere-datennutzung/ (letzter Zugriff: 17.2.2022).

23 Siehe z.B. die sogenannten „Priority Setting Partnerships“: www.jla.nihr.ac.uk/about-the-james-lind-alliance/about-psps.htm

24 Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, Tugwell P. 2012. Developing core outcome sets for clinical trials: issues to consider. Trials, 13: 132.

delt wurden, um ihre eigene Reputation fürchten.[25] Auch ohne zu wissen, welche konkreten Use & Access-Komitees bzw. Krankenhäuser entsprechende Anfragen negativ beschieden haben, ließe sich alleine dadurch, dass die betreffenden Fragestellungen veröffentlicht werden, evaluieren, ob besonders relevante Fragestellungen häufig nicht beforscht werden können.

Eine besondere Herausforderung für Sekundärnutzungen von Patientendaten stellen sogenannte Zufallsbefunde/Zufallsfunde da. Es ist möglich, dass bestimmte Auswertungsergebnisse im Kontext einer Sekundärnutzung für individuelle Patient:innen von so erheblicher Bedeutung sind, dass Ärzt:innen oder Forscher:innen eine Kontaktaufnahme als dringend notwendig erachten. Neben diesem akuten Bedarf an einer Kommunikation von Zufallsbefunden, können Auswertungsergebnisse auch nur ein möglicherweise gesundheitlich relevantes Informationspotential für individuelle Patient:innen haben. Auch in diesem Kontext sind deshalb Prozeduren und Best-Practice Standards u. a. zur Qualität/Diagnosesicherheit, zur Dringlichkeit der Kommunikation oder zur Qualität der Kommunikation von Zufallsbefunden nötig. Viele dieser Prozeduren werden nur dann Best-Practice-Standards entwickeln können, wenn die erwünschten und unerwünschten Effekte der Kommunikation von Zufallsbefunden nachverfolgt werden. Es besteht hier eine gewisse Ähnlichkeit zur Früherkennung von Erkrankungen oder zu Risikofaktoren bei asymptomatischen Personen. Aus der Krebsfrüherkennung sind z. B. erhebliche Herausforderungen im Umgang mit Fehl- wie auch Überdiagnosen bekannt.[26]

Diskussion

Die Sekundärnutzung von Patientendaten birgt hohe Potentiale für neue und wichtige Erkenntnisse sowohl im Kontext der Versorgungsforschung und Qualitätssicherung (Stichwort: Lernendes Gesundheitssystem) als auch für die Entwicklung innovativer Verfahren zur besseren Prävention und für individuell ausgerichtete Therapieempfehlungen (Stichwort: stratifizierte/personalisierte Medizin). Dieses Potential lässt sich jedoch nur heben, wenn die Ergebnisse von Sekundärnutzungen vertrauenswürdig und nützlich für die Wissenschaft und die Patient:innen sind. Damit Vertrauenswürdigkeit und Nützlichkeit nicht allein

25 Simon GE, Coronado G, DeBar LL, Dember LM, Green BB, Huang SS, Jarvik JG, Mor V, Ramsberg J, Septimus EJ et al. 2017. Data Sharing and Embedded Research. Ann Intern Med, 167(9): 668670.

26 Davies L, Petitti DB, Woo M, Lin JS. 2018. Defining, Estimating, and Communicating Overdiagnosis in Cancer Screening. Ann Intern Med, 169(11): 824.

theoretische Ideale bleiben, bedarf es der Konsensfindung über und der Implementierung konkreter Prozeduren und Best-Practice-Standards zur Förderung robuster, transparenter und patientenorientierter Sekundärnutzungen.

In diesem Beitrag wurde primär versucht, das Problembewusstsein für diesen zu großen Teilen noch ausstehenden Aufgabenbereich zu schärfen. Neben den Forschenden selber sind auch die systemverantwortlichen Institutionen wie Forschungseinrichtungen, aber auch Fachgesellschaften und Forschungsförderer gefragt, die Vertrauenswürdigkeit und Nützlichkeit von Sekundäranalysen so zu operationalisieren, dass sie umsetzbar und evaluierbar werden.

Bei vielen Themen wie der Registrierung und Ergebnisveröffentlichung kann man sich bei der Sekundärnutzung an bereits bestehenden Best-Practice-Standards aus dem Bereich der klinischen Forschung [27] orientieren und muss diese wahrscheinlich nur leicht spezifizieren. Bei anderen Themen werden stärkere Modifizierungen zu existierenden Standards oder genuin eigene Standards zu entwickeln sein. Dazu gehören Standards zur Bewertung der Datenqualität, der Reproduzierbarkeit von Analysetechniken mit maschinellem Lernen oder künstlicher Intelligenz, der Kommunikation von Zufallsfunden oder der Transparenz zu nicht genehmigten Sekundärnutzungen.

Ein wichtiger erster Schritt auf diesem Weg wäre ein Agenda- oder ein Roadmap-Prozess, der aufzeigt, wo Prozeduren für vertrauenswürdige und nützliche Sekundärnutzung benötigt werden, für welche dieser Prozeduren es bereits konsensfähige Standards gibt und für welche noch nicht. Eine solche Agenda würde das Problembewusstsein der Forschenden und der systemverantwortlichen Institutionen fördern und die Grundlage für eine partizipative, transparente und effiziente Entwicklung von noch ausstehenden Standards für die verantwortungsvolle Sekundärnutzung von Patientendaten schaffen.

27 von Niederhausern B, Guyatt GH, Briel M, Pauli-Magnus C. 2018. Academic response to improving value and reducing waste: A comprehensive framework for INcreasing QUality In patient-oriented academic clinical REsearch (INQUIRE). PLoS Med, 15(6): e1002580.

Anhang

Tabelle 1: Konzeptionelle und empirische Fragen (orientierend) zu Prozeduren für verantwortungsvolle Sekundärforschung

Prinzipien	Spezifizierung (orientierend)	Konzeptionelle Fragen: Prozeduren (orientierend)	Empirische Fragen: Qualitätssicherung / Meta-Research (orientierend)
Vertrauenswürdig	Datenqualität	Wie lässt sich die Qualität der für die Sekundärnutzung zur Verfügung stehenden Daten operationalisieren? Spezifisch für bestimmte Datentypen, Verwendungen etc.?	Wie gut ist die entsprechende Datenqualität?
	Analysen/ Ergebnis-Validität	Wie lässt sich die Robustheit der Analysen/Ergebnisse operationalisieren? Spezifisch für bestimmte Fragestellungen, Datentypen?	Wie robust sind die bestimmte Analysen/ Ergebnisse?
	Reproduzierbarkeit	Wie lässt sich die Reproduzierbarkeit operationalisieren?	Wie oft ist eine Reproduzierbarkeit in bestimmten Bereichen der Sekundärnutzung gegeben?
Nützlich-Transparent	Anfragen zur Sekundärnutzung	Welche Form von Transparenz bedarf es zu genehmigten und nicht genehmigten Anfragen?	Was sind häufige Gründe für nicht-genehmigte Anträge? Barrieren für besonders patientenorientierte Sekundärnutzung
	Methoden der Sekundärnutzung (Analysen/ Code)	Welche Form von Transparenz zu Studienprotokollen ist erforderlich?	Welche Methoden zur Verringerung von Verzerrungen/Bias werden wann und wie häufig verwendet?
	Ergebnisse der Sekundärnutzungen	Welche Form von Transparenz ist erforderlich?	Wie vollständig, zeitnah erfolgt die Ergebnisveröffentlichung? Wie konsistent sind die Ergebnisse mit dem Inhalt/den Methoden der Anfragen und Protokolle?

Nützlich-Relevant	Patienten-/Bürgerbeteiligung	Welche Form/Qualität von Patienten-/Bürgerbeteiligung bedarf es bei der Begutachtung von individuellen Projekten oder der übergreifenden Performance der Sekundärnutzung?	Werden Prozesse der Patienten-/Bürgerbeteiligung angemessen umgesetzt?
	Patientenorientierung	Wie sähe ein Prozess aus, um besonders patientenorientierte, versorgungsrelevante Sekundärnutzungen festzulegen?	Werden besonders relevante Fragen in Förderprogrammen adressiert?
	Zufallsfunde	Welcher Vorgaben bedarf es zur Kommunikation von Zufallsfunden? Welche Evaluation der Effekte von Zufallsfunden ist erforderlich?	Diagnostische Validität: Wie oft handelt es sich um falsche Befunde? Klinische Validität: Wie oft handelt es sich bei richtigen Befunden um Überdiagnosen/Übertherapie?

AUTOMATISIERTE ENTSCHEIDUNGEN: ASPEKTE VON FAIRNESS, DATENQUALITÄT UND PRIVACY

Frauke Kreuter, Christoph Kern, Patrick Oliver Schenk

Ohne Zweifel sind die Möglichkeiten des Einsatzes von künstlicher Intelligenz (KI) in der Medizin vielversprechend. Dieser Band gibt Anregungen und lässt uns hoffnungsvoll in die Zukunft blicken. Damit sich diese Zukunft so realisiert, wie wir uns das erhoffen, ist es wichtig, einige Aspekte der KI-Anwendungen zu betrachten, die außerhalb der reinen Anwendung von Algorithmen liegen. Dazu gehören Fairness, Datenqualität und Privacy.

Ein Wort vorab zur Verwendung der Begrifflichkeiten: KI, Algorithmen und maschinelles Lernen haben für Statistiker und Datenwissenschaftler ganz spezifische Bedeutungen. Im breiten Sprachgebrauch werden sie allerdings oft in einem Atemzug genannt und austauschbar verwendet.[1] Das liegt an einer Gemeinsamkeit, die für diesen Beitrag von besonderer Bedeutung ist: Sie alle extrahieren Informationen und generieren Vorhersagen aus Daten. Die Aspekte, die dieser Beitrag behandelt, sind relevant für alle Situationen, in denen Entscheidungen mit Hilfe von Modellen getroffen werden, die auf Daten beruhen.

Im Vergleich zu klassischen statistischen Verfahren wie Regression haben KI-Verfahren eigene Stärken: Sie können unter potentiell sehr vielen Prädiktoren die relevantesten automatisch herausfiltern (z. B. in der Genomik sehr wichtig); auch sehr komplexe Strukturen wie nichtlineare Zusammenhänge oder Interaktionen können flexibel erkannt werden; die durchschnittliche Güte von Prognosen ist im Allgemeinen höher. Während die meisten heutigen KI-Anwendungen nur ein *allgemein* gut passendes Modell entwickeln, liegen Hoffnungen auch im vermehrten Einsatz von Verfahren zur Erkennung heterogener Effekte zwischen Gruppen oder Individuen (z. B. Präzisionsmedizin). Zu den Nachteilen von KI gehören insbesondere eine teils viel geringere Interpretierbarkeit, Verständlichkeit und Transparenz sowie ein oft viel höherer Bedarf an Daten (Beobachtungen) und Computer-Ressourcen. Maße dafür, wie unsicher die Ergebnisse von KI sind, werden noch entwickelt bzw. sind heute oft überoptimistisch. Auch wird die o. g. höhere

1 Was heute mit dem Oberbegriff KI bezeichnet wird, ist sehr oft ausschließlich maschinelles Lernen. Eine Einführung in letzteres bieten James G., Witten D., Hastie T., & Tibshirani R. 2021. An Introduction to Statistical Learning, Springer: New York, 2. Auflage, Buch und Kurs verfügbar via www.statlearning.com.

mittlere Prognosegüte typischerweise erkauft mit einem höheren, systematischen Bias – verzerrte Ergebnisse, die z. B., aber nicht immer, v. a. auf bestimmte Gruppen entfallen. Maße für die Datenabhängigkeit fehlen weitgehend.

Wie klassische Statistik erkennt und repliziert KI Muster in Daten, und zwar unabhängig davon ob diese Muster nun auf Datenprobleme oder auf „echte", uns interessierende Zusammenhänge zurückgehen. Wohnen den Daten, anhand derer ein Modell trainiert wird, Probleme inne, ist nicht zu erwarten, dass KI diese von selbst erkennt und löst.[2]

Im ersten der drei Teile dieses Beitrags stellen wir einige KI-Anwendungen vor. Gemeinsam ist den Anwendungen, dass sie Daten aus unterschiedlichen Quellen verwenden und Vorhersagen für unterschiedlichste Gruppen treffen. Gemeinsam ist ihnen auch, dass sie eine Balance zwischen der Qualität der Vorhersage insgesamt und der Vorhersagegüte für einzelne Gruppen finden müssten. Werden die gruppenspezifischen Vorhersagen nicht berücksichtigt, so kommt es leicht zu unintendierten Ungleichbehandlungen einzelner Gruppen. Mittlerweile steht eine Reihe von Metriken zur Verfügung, um Ergebnisse von KI-Anwendungen auf dieses Problem hin zu prüfen. Ob die Algorithmen sich fair verhalten, ist hierbei eine zentrale Frage.

Die mögliche Ungleichbehandlung kann, muss aber nicht unbedingt ein Problem sein. Denn aus soziologischer Sicht kann eine ungleiche Behandlung verschiedener Bevölkerungsgruppen durchaus intendiert sein. Dies werden wir im zweiten Teil diskutieren. Das Problem der Fairness wird sich nicht rein technisch lösen lassen, sondern bedarf einer gesellschaftlichen Diskussion über allgemeine Gerechtigkeitsprinzipien. Aber auch wenn sich eine Gesellschaft darauf einigt, welche Gerechtigkeitsprinzipien angewendet werden sollen, können die KI-Anwendungen nur so gut sein, wie die Daten, die in sie hineinfließen.

Ein Problem, das sich bei der Sammlung oder Generierung von Daten immer stellt, ist das systematische Fehlen von Informationen. Dies kann verschiedenste Ursachen haben; eine, an der wir etwas ändern können, ist die Frage des Datenzugangs. Wir sind in Deutschland aus gutem Grund darauf bedacht, den Datenschutz hochzuhalten. Ohne Zweifel ist die Selbstbestimmung darüber, welche

2 Ein einfaches Mehr der gleichen Daten löst das Problem nicht. Fehlt eine gesellschaftliche Gruppe völlig oder hat eine Variable einen systematischen Messfehler, ändert sich allein durch eine höhere Beobachtungszahl nichts an der fehlenden Repräsentanz oder dem Messfehler.

Informationen von einem und über einen verwendet werden können, ein hohes Gut. Es ist jedoch für die oder den Einzelne/-n eine schwer zu bewältigende Aufgabe, in jeder Situation zu entscheiden, ob eigene Daten freigegeben werden sollten oder nicht. Die daraus resultierenden Konsequenzen zu überblicken ist nicht nur schwierig, sondern oft unmöglich. Wir regen im dritten Teil deshalb an, einen normativen Blick auf diese Frage zu werfen und zu überlegen, in welchem Kontext die Nutzung welcher Daten legitim ist. Anstatt die Datennutzung über Datentypen zu regeln, könnte es sich lohnen, mehr darauf zu achten, mit welchem Ergebnis Daten genutzt werden. Technisch stehen bereits viele Lösungen zur Verfügung, die es ermöglichen würden, Daten für die Hebung des Allgemeinwohls besser zu nutzen. Dieser Aufgabe sollten wir uns stellen.

AUTOMATISIERTE ENTSCHEIDUNGEN IM ÖFFENTLICHEN SEKTOR

Häufig steht das Ergebnis eines Algorithmus nicht für sich, sondern ist nur der erste Schritt: als Basis von Entscheidungen (zweiter Schritt). Überall dort, wo knappe Ressourcen eingesetzt werden, besteht das Bedürfnis, den Einsatz dieser Ressourcen so zu optimieren, dass das Ergebnis möglichst effizient ist. Datenbasierte Vorhersagen über z. B. den Bedarf einer Zuteilung oder den möglichen Ausgang einer Handlung können beim Einsatz knapper Ressourcen helfen. Sind x Einheiten einer Ressource auf mehr als x Personen zu verteilen, kann ein Algorithmus z. B. diejenigen x Individuen mit dem größten Bedarf ermitteln. Koppelt man die Entscheidung dann fest an das Ergebnis des Algorithmus (die x Personen mit dem größten Bedarf werden bedacht, alle anderen nicht), d. h. automatisiert man die Entscheidung, spricht man von Automated Decision Making (ADM).[3] Dies ist vor allem dann attraktiv, wenn sich Ereignisse stabil aus vorhandenen Daten vorhersagen lassen, also wenn (und besser: nur, dann wenn) sich immer gleiche Muster in Daten finden lassen, die verlässlich auf ein Problem oder ein Ereignis hinweisen. Ein Vorteil der Automatisierung von Entscheidungen liegt in der Beschleunigung von Prozessen. Ebenso erhofft man, dass durch die feste Anbindung an das Ergebnis des Algorithmus nur noch relevante Größen einen Einfluss auf die Entscheidung haben und diese objektiver als durch Menschen gefällte Entscheidungen sind.

3 Die Übergänge dazu, dass die Verknüpfung nicht eins-zu-eins ist, sondern noch Entscheidungsspielraum besteht (Semi-automated Decision Making), ist durchaus fließend. Die klassische Situation hingegen kennzeichnet, dass die Resultate eines Algorithmus nur eine Information unter vielen für (einen) menschliche(n) Entscheider darstellen.

ADM im Gesundheitswesen

Wie sehen automatisierte Entscheidungen im Gesundheitswesen aus? In den USA nutzen Krankenversicherungen Algorithmen, um das zukünftige Krankheitsrisiko und damit den Bedarf für einen besonders hohen Versorgungsaufwand abzuschätzen. Diejenigen mit einem hohen Risiko werden dann mit besonderen Vorsorgeprogrammen unterstützt.[4] In Israel nutzen Ärzt:innen Vorhersagemodelle, um zu entscheiden, welche Patienten welche personalisierten Behandlungen bekommen sollen.[5] In Chicago werden Verfahren des maschinellen Lernens genutzt um vorherzusagen, welche HIV-Positiven wahrscheinlich aus der kontinuierlichen Behandlung und Betreuung herausfallen und damit sich selbst und andere gefährden könnten.[6] Im US-Bundesstaat Kansas wird versucht, mit Hilfe von KI die Spirale von unbehandelten psychiatrischen Erkrankungen und Verhaftungen zu durchbrechen. Personen, die laut Modell eine hohe Wahrscheinlichkeit haben, straffällig zu werden, werden je nach vorhandenen Ressourcen priorisiert behandelt.[7] Hinter diesen Einsätzen der Algorithmen verbirgt sich die Hoffnung, die begrenzten Vorhersagefähigkeiten von Menschen mit statistischen, algorithmischen Vorhersagetechniken zu verbessern oder gar zu ersetzen.[8]

Was kann da schiefgehen?

Der Einsatz von Algorithmen, der im amerikanischen Justizsystem eingeführt wurde, um Entscheidungen effizienter zu gestalten und der Subjektivität von Richtern vorzubeugen, hat jedoch zu Ernüchterung geführt. In Untersuchungen der gemeinnützigen Nachrichtenredaktion ProPublica, zeigte sich, dass nur 20 Prozent der Personen, für die Gewaltverbrechen vorhergesagt worden waren, diese auch tatsächlich begingen. Bei der Vorhersage, wer wieder straffällig werden

4 Obermeyer Z, Powers B, Vogeli C, & Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464): 447–453.

5 Dagan N, Cohen-Stavi CJ, Avgil Tsadok M. et al. 2019. Translating clinical trial results into personalized recommendations by considering multiple outcomes and subjective views.npj Digital Medicine 2(81).

6 Ramachandran A, Kumar A, Koenig H. et al. 2020. Predictive Analytics for Retention in Care in an Urban HIV Clinic. Scientific Reports 10(6421).

7 Rodolfa KT, Lamba H & Ghani R. 2021. Empirical observation of negligible fairness–accuracy tradeoffs in machine learning for public policy. Nature Machine Intelligence 3: 896–904.

8 Engelmann J & Puntschuh M. 2020. KI im Behördeneinsatz: Erfahrungen und Empfehlungen. Fraunhofer-Institut für Offene Kommunikationssysteme, Berlin. http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-6350714.pdf

würde, machte der Algorithmus bei schwarzen und weißen Angeklagten in etwa gleichem Maße, aber auf sehr unterschiedliche Weise Fehler. Der Algorithmus stufte schwarze Angeklagte fast doppelt so häufig fälschlicherweise als rückfällig ein wie weiße Angeklagte (falsch-positiv), während weiße Angeklagte wesentlich häufiger fälschlicherweise als nicht-rückfällig eingestuft wurden (falsch-negativ).[9]

Seit ProPublica auf diese gruppenspezifischen Vorhersageunterschiede aufmerksam gemacht hat, überschlägt sich die KI-Forschung mit der Entwicklung und dem Einsatz von Maßzahlen bzw. Metriken zur Bestimmung von ‚Fairness'. Zugleich schreiben Regularien vor, dass in Antidiskriminierungsgesetzen definierte, geschützte Merkmale (wie Geschlecht, Alter, Herkunft) nicht zum Training von KI-Algorithmen verwendet und Entscheidungen nicht allein auf Algorithmen basieren dürfen.

Weder Regularien noch Maßzahlen haben bisher zum gewünschten Erfolg geführt.

- Selbst wenn geschützte Merkmale nicht für das Modelltraining verwendet werden, können die Vorhersagen unterschiedlich gut für geschützte Gruppen sein.[10] Wenn geschützte Merkmale komplett aus den Daten entfernt werden, ist es umgekehrt sehr schwer, überhaupt zu evaluieren, ob eine bestimmte Gruppe systematisch benachteiligt wird.

- Selbst wenn menschliche Entscheider den Vorschlag des Algorithmus lediglich als eine Informationsquelle hinzuziehen sollen, werden sie doch stark davon beeinflusst und neigen dazu, Vorgeschlagenes zu bestätigen.[11]

- Selbst wenn Metriken zur Überprüfung der ‚Fairness' von Algorithmen zur Verfügung stehen, leiden sie an dem inhärenten Problem, dass sie Zielgrößen formulieren, die von Algorithmen nicht alle gleichzeitig eingehalten werden können.[12]

9 Angwin J, Larson J, Mattu S, Kirchner L. (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

10 Pope DG, and Sydnor JR. 2011. Implementing Anti-discrimination Policies in Statistical Profiling Models. American Economic Journal: Economic Policy 3(3): 206–231.

11 Goddard K, Roudsari A, & Wyatt JC. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19(1): 121–127.

12 Berk R, Heidari H, Jabbari S, Kearns M, & Roth A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research 50(1): 3–44.

Der letzte Punkt ergibt sich daraus, dass wir zum einen in der Praxis in den allermeisten Situationen unterschiedliche Basisraten für verschiedene gesellschaftliche Gruppen vorfinden. Zum Beispiel geht man davon aus, dass Frauen eher zu Depressionen neigen oder dass junge Männer für die überwiegende Mehrheit der Gewaltverbrechen verantwortlich sind. Zum anderen sind Modelle nie perfekt und die Modellgüte unterscheidet sich oft zwischen den Gruppen. Derartige Unterschiede wirken sich auf die statistisch erreichbare Fairness aus, d. h. auf die Frage, welche Metriken gleichzeitig erfüllt werden können. Abzuwägen gilt z. B., ob man einen Anstieg von falsch-negativen Vorhersagen in Kauf nimmt, um mehr Fairness zwischen den Gruppen herzustellen.[13]

Auch wenn ethische Prinzipien in der Medizin eine optimale Behandlung aller gebieten, gibt es auch hier zahlreiche Situationen, in denen aufgrund bestehender Ressourcenknappheit (feste Gesamtbudgets, fixe Kapazitäten oder Wartezeiten) Entscheidungen über Behandlungsprioritäten oder die Zuteilung bestimmter Vorsorgemaßnahmen getroffen werden müssen. Derzeit werden in Deutschland Leitlinien der Fachgesellschaften in diesen Entscheidungssituationen verwendet, individuelle Interpretationen und Abwägungen sind in der Regel aber unvermeidbar. Wenn KI systematisch eingesetzt werden soll, wird eine Diskussion über Leitlinien und Interpretationen umso wichtiger. Denkbar ist, dass beim Einsatz von KI bestimmte systematische Verzerrungen in ganz anderer Weise skalieren,[14] individuelle Biases eines einzelnen Entscheiders jedoch überschrieben werden. Das heißt, dass anstatt eine Reihe Entscheider mit unterschiedlichen Biases zu haben, wird beim Einsatz von KI für alle Entscheidungen der Bias der Mehrheit wirksam. Im oben erwähnten Gerichtskontext kann man sich das wie folgt vorstellen: Im Idealfall skaliert man die beste Richterin. Im schlimmsten Fall sind alle richterlichen Entscheidungen von einem rassistischen Bias geprägt.

13 Ebd.

14 Gerade bei automatisierten Entscheidungen ist es möglich, dass einmal eingeschriebene Verzerrungen sich selbst bestätigen oder gar verstärken: Die aufgrund des Bias in den Ergebnissen des Algorithmus verzerrten Entscheidungen produzieren neue Daten, welche ebenso oder gar stärker verzerrt sind – und wieder in den Algorithmus eingehen.

PRINZIPIEN DER VERTEILUNGSGERECHTIGKEIT

Wenn Entscheidungen zum gezielten Einsatz von knappen Ressourcen getroffen werden, stellt sich schnell die Frage, was gerecht oder was fair ist. Die Fairnessmetriken der KI-Forschung sehen eine Allokation von Ressourcen in der Regel dann als fair an, wenn Personen, die sich durch bestimmte geschützte Attribute unterscheiden (Geschlecht, ethnische Zugehörigkeit etc.), identische Entscheidungen mit identischer Fehlerwahrscheinlichkeit zugewiesen werden. Schaut man aus einer sozialwissenschaftlichen Perspektive auf diese Metriken, wird schnell deutlich, dass prominente Fairnessmetriken der KI-Forschung nicht gut mit den verschiedenen Ansätzen der Theorien zur Verteilungsgerechtigkeit übereinstimmen.

Nachstehende Tabelle[15] zeigt eine Zusammenfassung von zentralen Gerechtigkeitsprinzipien und eine Illustration ihrer Anwendung auf das Beispiel der HIV-Prävention. Vorhersagemetriken (d. h. Metriken zum Vergleich von Vorhersagefehlern) wenden Chancengleichheit (equality of opportunity) auf die Verteilung von Vorhersagefehlern an. Das heißt, geschützte Personengruppen sollen die gleiche Chance auf eine nicht-fehlerhafte Vorhersage haben.

Entscheidungsmetriken (d. h. Metriken zum Vergleich der Eintrittswahrscheinlichkeiten des vorhergesagten Ereignisses) implementieren Vorstellungen von Egalitarismus. Hierbei sollen Personen nicht für bestimmte Merkmale (z. B. für ihr Geschlecht) verantwortlich gemacht werden und sollten deshalb die gleichen Risikovorhersagen erhalten. Wir argumentieren, dass diese beiden Verbindungen unzureichend sind. Gerechtigkeitsprinzipien wie Equality, Desert, Need und Efficiency werden durch bestehende Fairnessmetriken nicht ausreichend erfasst. (Die englischen Fachbegriffe werden in der folgenden Tabelle erläutert.)

15 Ausführlich behandelt in Kuppler M, Kern C, Bach RL & Kreuter F. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? https://arxiv.org/abs/2105.01441

Tabelle 1: Ausgewählte Prinzipien der Verteilungsgerechtigkeit und deren Illustration am Beispiel der Verteilung von Präventionsmaßnahmen für HIV-Infizierte. Adaptiert nach Kuppler u. a. 2021 in https://arxiv.org/abs/2105.01441

Gerechtigkeitsprinzip (engl. terms)	**Verteilungsregel**	**Beispiel**
Equality	Verteile R Ressourcen auf Individuum X genau dann, wenn die für X vorhandenen Ressourcen Y so sind, dass die Hinzugabe von R die gesamtgesellschaftliche Ungleichheit minimiert.	Verteile Präventionsmaßnahmen gegen Versorgungsausfall für HIV-Infizierte so, dass alle HIV-Positiven gleich gut versorgt sind.
Desert	Weise R Ressourcenanteile Individuum X genau dann zu, wenn diese auf X zutreffen.	Lasse Präventionsmaßnahmen denen zukommen, die ihre Beiträge zur Krankenversicherung bezahlt haben.
Need	Weise R Ressourcenanteile Individuum X genau dann zu, wenn X diese braucht.	Fokussiere Präventionsmaßnahmen auf diejenigen, deren Einkommen unter eine bestimmte Schwelle fällt.
Efficiency	Verteile R Ressourcen auf Individuum X genau dann, wenn die Hinzugabe von R den gesamtgesellschaftlichen Nutzen maximiert.	Fokussiere Präventionsmaßnahmen auf die am stärksten von den Maßnahmen profitierende Gruppe.
Equality of Opportunity	Weise allen X mit gleichen Attributen die gleiche Anzahl an R Ressourcenanteilen zu.	Stelle sicher, dass alle Gruppen von HIV-Infizierten die gleiche Chance haben, unterstützt zu werden.

Ein Ansatzpunkt wäre, die ADM-Prozesse in zwei Schritten zu denken: einen Vorhersage- und einen Entscheidungsschritt.[16] Das Ziel des Vorhersageschritts besteht darin, ein möglichst akkurates Modell davon zu erstellen, wie die Welt tatsächlich ist bzw. wie sie mit den gegebenen Daten abgebildet werden kann. Für den Entscheidungsschritt sollte dann explizit eine Verteilungsregel definiert werden. Diese

16 Kuppler, M, Kern, C, Bach, RL, Kreuter, F. 2022. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. Frontiers in Sociology. https://doi.org/10.3389/fsoc.2022.883999

resultiert nicht aus einem Modell, sondern aus einer gesellschaftlichen Diskussion über Werte und erwünschte Zielzustände. Die Korrektur von gesellschaftlichen Verzerrungen ist Aufgabe der Verteilungsregeln, nicht die des Vorhersagemodells. Auch heute findet dieser Entscheidungsschritt schon statt, auch ohne den Einsatz von KI. Im Kontext des Einsatzes von KI besteht allerdings die erhöhte Gefahr, dass Verzerrungen im Vorhersageschritt für die menschlichen Entscheider[17] schwer erkennbar sind und systematisch zu dem Entscheidungsschritt durchgereicht werden. Auch bei dem expliziten Einsatz von Menschen im zweiten Schritt, ist ein derartiges Vorgehen wahrscheinlich, da Menschen doch dazu neigen, vorgefertigte Annahmen[18] bzw. Vorgeschlagenes zu bestätigen.[19]

DIE ROLLE DATENGENERIERENDER PROZESSE

Wie gut ein Vorhersagemodell die Welt abbilden kann, hängt entscheidend davon ab, welche Daten zum Training der Algorithmen verwendet werden. In der Praxis beobachten wir häufig eine Diskrepanz zwischen der Art und Weise, wie die Welt tatsächlich ist und wie die Welt in den Trainingsdaten von KI-Modellen abgebildet wird. Statistische Verzerrungen entstehen z. B. durch Repräsentation/Sampling Bias (d. h. die Daten repräsentieren nicht die Gesamtpopulation, auf die das Vorhersagemodell später angewendet werden soll) oder Messfehler (d. h. es gibt systematische Diskrepanzen zwischen gemessenen und wahren Attributen).

Statistische Verzerrungen können eine robuste Anwendung von Vorhersagemodellen in der Medizin stark einschränken. Je nach Vollständigkeit der dem Training der Modelle zugrunde liegenden Daten kann die Übereinstimmung zwischen vorhergesagten und beobachteten (Krankheits-) Risiken für bestimmte Bevölkerungsgruppen sehr unterschiedlich sein.[20] So zeigen Barda et al. (2021)[21]

17 Sofern solche noch eingebunden sind. Ist die Entscheidung absolut automatisiert, entfällt die Last vollends auf die Entwickler und ggf. Prüfer (audits) des Algorithmus.

18 Klayman J. 1995. Varieties of Confirmation Bias, in: Busemeyer J, Hastie R, & Medin D L (Eds.), Psychology of Learning and Motivation, Vol. 32: 385–418. Academic Press.

19 Siehe die Ausführungen von Tourangeau R et al. 2000. The Psychology of Survey Response, Cambridge University Press, zu Acquiescence Bias.

20 Im Allgemeinen fehlen zudem auch noch belastbare Maße für die Unsicherheit der jeweiligen Prognose.

21 Barda N, Yona G, Rothblum GN, Greenland P, Leibowitz M, Balicer R, Bachmat E, & Dagan N. 2021. Addressing bias in prediction models by improving subpopulation calibration. Journal of the American Medical Informatics Association 28(3): 549–558.

beispielsweise für ein Modell zur Einschätzung des Risikos von Herz-Kreislauf-Erkrankungen eindrucksvoll, zu welchen Verzerrungen und Vorhersagefehlern es für Teilpopulationen kommen kann. Solche Effekte können beispielsweise dann auftreten, wenn Teilgruppen in der Modellentwicklung unterrepräsentiert sind. Das ist ein Problem jedes datenbasierten Modells, das wir auch heute schon kennen und das bei KI-Anwendungen ebenso besonderer Aufmerksamkeit bedarf.[22]

Von wem haben wir welche Daten?

Vor der Anwendung eines KI-Algorithmus sollte man sich deshalb immer fragen, von welchen Populationen Informationen während der Modelltrainingsphase zur Verfügung standen. Insbesondere sollten die Personengruppen in den Trainingsdaten den Personengruppen entsprechen, für die Vorhersagen gemacht werden.[23] Nicht alle Datenquellen umfassen alle Personengruppen, für die Vorhersagen gemacht werden sollen. So gibt es beispielsweise sozial-strukturelle Unterschiede im Hinblick auf den Besitz und die Nutzung von elektronischen Geräten wie z. B. Smartphones, welche zur Generierung von Trainingsdaten genutzt werden können.[24] Um Unzulänglichkeiten einzelner Datensätze zu begegnen, werden deshalb zunehmend Daten aus verschiedensten Quellen miteinander verknüpft.[25] Allerdings kann auch diese Verknüpfung für Teilgruppen systematisch häufiger fehlschlagen. In Deutschland und Europa wird an vielen Stellen verlangt, dass Personen explizit einer Weitergabe von Daten zustimmen. Leider wird die Zustimmung oft nicht basierend auf inhaltlichen Überlegungen gegeben oder verweigert, sondern erfolgt situativ und oft ohne fest verankerte Einstellungen. Dies führt dazu, dass Personen, die nicht nachvollziehen können, was mit ihren Daten passiert, oder die wenig Vertrauen in daten-erhebende Institutionen haben,

22 Ein weiterer Grund sind Messfehler, die sich systematisch zwischen Gruppen unterscheiden. Lernt ein Algorithmus also z. B. nicht anhand des wahren Krankheitsstatus, sondern anhand von gestellten Diagnosen, die für manche Teilgruppen seltener richtig sind, so wird auch ein Algorithmus (und darauf aufbauendes ADM) diese Verzerrungen replizieren; er erkennt nicht von selbst, wenn die ihm gegebenen Informationen falsch sind.

23 Ein bekanntes Problem in der Medikamentenentwicklung, wenn z. B. (schwangere) Frauen nicht in die Studienphasen eingebunden werden, Medikamente später aber auch für sie genutzt werden.

24 Keusch F, Bähr S, Haas GC., Kreuter F, Trappmann M. 2020. Coverage Error in Data Collection Combining Mobile Surveys with Passive Measurement Using Apps: Data from a German National Survey. Sociological Methods & Research, https://doi.org/10.1177/0049124120914924

25 Die Nutzung neuer Variablen und der hohe Bedarf an Beobachtungen von KI sind weitere Gründe.

weniger häufig zustimmen.[26] Dadurch stehen für diese Personen dann später weit weniger Trainingsdaten zur Verfügung.

Wir konnten in verschiedenen Studien zeigen, dass die Zustimmungsrate zu Verknüpfungen besonders empfindlich auf bestimmte Designmerkmale reagiert, z. B. darauf, wo die Zustimmungsfrage im Fragebogen platziert ist und wie die Frage formuliert ist.[27] Für uns deutet dies darauf hin, dass die Einstellung zur Verknüpfung nicht so stark ausgeprägt ist, wie es die Vorschriften, die eine solche Zustimmung vorschreiben, vermuten lassen. Möglicherweise müssen wir uns aber auch einfach davon verabschieden, dass Individuen (in jeder Situation neu) darüber entscheiden können und sollen, welche Daten weitergegeben oder verknüpft werden sollen. Die immense Häufigkeit solcher Situationen und die Unmöglichkeit, alle zukünftig möglichen Nutzungen zu bedenken, spräche ebenfalls dafür. Eine Alternative wäre hier, stärker gesellschaftlich über angemessene Datenströme nachzudenken.

Welche Daten dürfen wir wann nutzen?

Helen Nissenbaum, Philosophin und Professorin für Informationswissenschaften der Cornell Tech University in den USA, befasst sich seit Jahren mit der Angemessenheit von Datenströmen. Unter dem Stichwort „contextual integrity" (CI) definiert sie Bedingungen, unter denen eine Datenverarbeitungspraxis angemessen ist. CI besagt, dass Datenübertragungen den Erwartungen an den Schutz der Privatsphäre entsprechen, wenn sie mit den Datenschutznormen übereinstimmen, die wiederum abhängig von der Art und den Umständen der gesammelten Informationen sowie den beteiligten Akteuren sind. Kontextuelle Informationsnormen geben für den Informationsfluss fünf Schlüsselparameter vor: (1) den Absender der Information, (2) den Empfänger der Information, (3) das Attribut oder die Art der Information, (4) das Subjekt der Information und (5) ein Übertragungsprinzip mit den Bedingungen für einen angemessenen Informationsfluss.[28] So ist es beispielsweise im Gesundheitswesen angemessen, dass Patienten (Sender und Subjekt) ihren Ärzten (Empfänger) Gesundheitsinformationen (Attribute) vertraulich zur Verfügung

26 Auch hier kann es zu systematischen Gruppenunterschieden kommen, wenn Bildung ein maßgeblicher Treiber des Verständnisses der Datennutzung ist, oder bestimmte soziale Gruppen aufgrund von Diskriminierung in der Vergangenheit vorsichtiger sind.

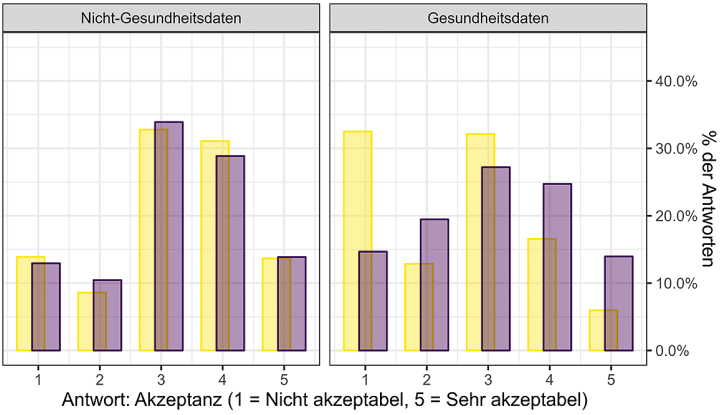
27 Sakshaug JW, Schmucker A, Kreuter F, Couper MP, Singer E. 2019. The Effect of Framing and Placement on Linkage Consent, Public Opinion Quarterly 83(S1): 289–308.

28 Nissenbaum H. 2010. Privacy in context: Technology, policy, and the integrity of social life. Stanford, Calif.: Stanford Law Books.

stellen (Übertragungsprinzip). Dieser Ablauf ist normenkonform und unproblematisch. Leitet eine Arztpraxis jedoch medizinische Informationen an einen anderen Empfänger, z. B. den Arbeitgeber eines Patienten, weiter, so ist das Prinzip der vertraulichen Übertragung verletzt. Bei der Beurteilung der Angemessenheit müssen immer alle Parameter berücksichtigt werden. Eine Datenschutzregel, die sich nur auf die Attribute oder Arten der Informationen bezieht, ist unzureichend. Die Wahrnehmung der Bevölkerung, welche Nutzung akzeptabel ist, kann sich situativ durchaus verändern. Wie die Daten von rund 600 Befragten von Gerdon et al. (2021)[29] in der nebenstehenden Abbildung zeigen, war die Bereitschaft, eigene Daten zur Bekämpfung einer Pandemie zur Verfügung zu stellen, vor der COVID-19-Pandemie deutlich niedriger als nach Beginn. Diesen Unterschied sahen wir bei anderen Datentypen und anderen Nutzungen nicht.

Heißt das, jeder kann an die Daten ran?

Aus den Prinzipien der CI können auch Regeln für die Datenverarbeitungspraxis abgeleitet werden. Das sogenannte „Five Safes Framework" besteht aus einer Reihe von Grundsätzen, die es Datendiensten ermöglichen, einen sicheren Forschungszugang zu Daten anzubieten.



Siehe auch: Gerdon, Nissenbaum, Bach, Kreuter, Zins 2021. Harvard Data Science Review. https://doi.org/10.1162/99608f92.edf2fc97

29 Gerdon F, Nissenbaum H, Bach RL, Kreuter F, & Zins S. 2021. Individual Acceptance of Using Health Data for Private and Public Benefit: Changes During the COVID-19 Pandemic, Harvard Data Science Review : HDSR, 3(Spec. Iss. 1), 1–27, https://doi.org/10.1162/99608f92.edf2fc97

Das britische Office of National Statistics und einige andere Datenanbieter arbeiten seit den 2010er Jahren an diesem Rahmenwerk mit.[30] Die Five Safes haben sich als Best Practice für den Datenschutz etabliert und erfüllen gleichzeitig die Anforderungen an offene Wissenschaft und an Transparenz. Auch in Deutschland nutzen bereits einige Forschungsdatenzentren diese Prinzipien und erlauben Forschenden einen kontrollierten Zugang zu sensiblen oder vertraulichen Daten, so dass sie auf sichere und verantwortungsvolle Weise auf Datensätze zugreifen und diese nutzen können.

- **Sichere Daten:** Die Daten werden so behandelt, dass die Vertraulichkeit gewahrt bleibt.

- **Sichere Projekte:** Forschungsprojekte werden auf ihre Angemessenheit geprüft.

- **Sichere Personen:** Forscher:innen sind geschult und autorisiert, Daten sicher zu nutzen.

- **Sichere Einstellungen:** Eine sichere Umgebung verhindert die unbefugte Nutzung.

- **Sichere Ergebnisse:** Geprüfte und genehmigte Ergebnisse, die nicht vertraulich sind.

In den USA nutzt die „Administrative Data Research Facility" (ADRF) der Coleridge Initiative eine sichere Umgebung innerhalb der Amazon AWS GovCloud zum Hosten vertraulicher Daten. Sie wurde vom U.S. Census Bureau eingerichtet, um die Nutzung von administrativen und anderweitig sensitiven personenbezogenen Daten zu erleichtern. Die ADRF folgt ebenfalls dem Rahmenwerk der Five Safes zum Schutz von Daten und hat bereits über 100 vertrauliche Datensätze von Bundes-, Landes- und Kommunalbehörden sowie von akademisch Forschenden in die Cloud-Umgebung eingespeist. Datenanbieter können ihre Daten in dieser Umgebung bereitstellen und den Zugriff und die Nutzung mit einer speziell dafür entwickelten App steuern und verfolgen. Eine breiter aufgestellte Verfügbarkeit von (hochwertigen) Daten hat auch den Vorteil, dass sich Sampling- oder andere

30 Ein Beispiel der Umsetzung in Australien finden sich hier https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework

Datenerhebungsstrategien, die systematisch Teile der Bevölkerung ausschließen (z. B. Kranke an der Teilnahme von Umfragen[31] oder Ältere bei der Teilnahme an Datenspenden[32]), weniger leicht auf die Modellentwicklung durchschlagen.

FAZIT

- KI hat ein hohes Potenzial zur Effizienzsteigerung und globalen Verbesserung von Entscheidungen. Unintendierte Konsequenzen sind denkbar und insbesondere bei naiver Anwendung durchaus wahrscheinlich.

- Metriken zur Beurteilung von Fairness sind wichtig; wichtiger ist jedoch, gesellschaftliche Klarheit darüber zu bekommen, welche Verteilungsgerechtigkeit angestrebt werden soll.

- Auch bei klaren Verteilungszielen kann variierende Datenqualität zu Fehlern bei dem der Verteilungsentscheidung zugrundeliegenden Vorhersagemodell führen.

- Fehlende Daten können oft durch die Verknüpfung verschiedenster Datenquellen ausgeglichen werden. Verlinkung bedarf häufig einer informierten Zustimmung.

- Informierter Zustimmung liegt in der Praxis oft nicht tatsächlich Informiertheit zugrunde. Experimentell lassen sich Zustimmungsraten leicht manipulieren. Zustimmungen sollten nicht als akkurate Abbildung der Einstellung der Betroffenen interpretiert werden.

- Contextual integrity (Kontext der Datenerhebung und der angemessene Fluss von Informationen) ist eine ernstzunehmende Alternative zur informierten Einwilligung und könnte Kernprinzip des Datenschutzes werden.

- Prinzipien des angemessenen Informationsflusses lassen sich praktisch in Konzepten sicherer Datennutzung (wie die Five Safes) implementieren.

31 Schnell R & Trappmann M. 2006. Konsequenzen der Panelmortalität im SOEP für Schätzungen der Lebenserwartung. https://www.uni-due.de/~hq0215/documents/schnell_tote_100306.pdf (27.3.2022).

32 Schnell R & Smid M. 2020. Methodological Problems and Solutions in Sampling for Epidemiological COVID-19 Research. Survey Research Methods, 14(2): 123–129.

AUTORINNEN UND AUTOREN

Barth, Rico (rico.barth@cape-it.de), Open Source Business Alliance, Chemnitz

Dössel, Olaf (olaf.doessel@kit.edu), Mitglied der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und Leiter des Instituts für Biomedizinische Technik, Karlsruhe Institut für Technologie (KIT)

Fröhlich, Holger (holger.froehlich@scai-fraunhofer.de), Universität Bonn, Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS)

Ganten, Peter (ganten@osb-allianz.de), Vorsitzender Open Source Business Alliance – Bundesverband Digitale Souveränität e. V., CEO Univention GmbH

Haufe, Stefan (haufe@tu-berlin.de), Technische Universität Berlin

Intemann, Timm (intemann@leibniz-bips.de), Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS), Bremen

Kern, Christoph (christoph.kern@stat.uni-muenchen.de), Ludwig-Maximilians-Universität München

Kreuter, Frauke (frauke.kreuter@stat.uni-muenchen.de), Ludwig-Maximilians-Universität München | University of Maryland

Lippert, Christoph (Christoph.Lippert@hpi.de), Hasso-Plattner-Institut, Potsdam

Pigeot, Iris (pigeot@leibniz-bips.de), Nationale Forschungsdateninfrastruktur für Gesundheitsdaten (NFDI4Health) und Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS) und Nationale Forschungsdateninfrastruktur für Gesundheitsdaten (NFDI4Health)), Bremen

Prause, Guido (guido.prause@mevis.fraunhofer.de), Fraunhofer-Institut für Digitale Medizin MEVIS, Bremen

Rutert, Britta (rutert@bbaw.de), bis Juli 2022 wissenschaftliche Mitarbeiterin der Interdisziplinären Arbeitsgruppe „Zukunft der Medizin: Gesundheit für alle“ der BBAW

Schäffter, Tobias (tobias.schaeffter@ptb.de), Mitglied der BBAW und Leiter der Physikalisch-Technischen Bundesanstalt (PTB), Berlin

Schenk, Patrick Oliver (patrick.schenk@stat.uni-muenchen.de), Ludwig-Maximilians-Universität München

Schöck, Fabian (fabian.schoeck@siemens-healthineers.com), Siemens Healthineers, Erlangen

Schwabe, Daniel (daniel.schwabe@ptb.de), Physikalisch-Technische Bundesanstalt, Berlin

Strech, Daniel (daniel.strech@charite.de), Charité – Universitätsmedizin Berlin

Urban, Manuela (urban@osb-alliance.com), Open Source Business Alliance, Chemnitz

Wright, Marvin N. (wright@leibniz-bips.de), Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS), Bremen

In der Reihe „Denkanstöße aus der Akademie“ erschienen bisher

1 / Nov 2015
Franz-Xaver Kaufmann, Hans Günter Hockerts, Stephan Leibfried, Michael Stolleis, Michael Zürn
Zur Entwicklung von Forschung und Lehre zur Sozialpolitik an Universitäten in der Bundesrepublik Deutschland (nur online)

2 / Dez 2018
Christoph Markschies
Zwei Texte zur Akademie der Wissenschaften im einundzwanzigsten Jahrhundert (nur online)

3 / März 2020
Carola Lentz, Andrea Noll
Wissenschaftskooperationen mit dem globalen Süden: Herausforderungen, Potentiale und Zukunftsvisionen (nur online)

4 / März 2021
Jochen Gläser, Wolf-Hagen Krauth, Christine Windbichler, Michael Zürn
Befangenheit und Expertise in Berufungsverfahren: Ein wissenschaftspolitischer Denkanstoß (online und gedruckt)

5 / Juni 2021
Andreas Radbruch, Konrad Reinhart (Hrsg.)
Nachhaltige Medizin (online und gedruckt)

6 / Juni 2021
Jutta Allmendinger, Martin Mann, Lukas Haffert, Christoph Markschies
Junge Wissenschaftler:innen und die Pandemie: Unterstützung und systematische Verbesserungen – in der Krise und über die Krise hinaus (nur online)

7 / Nov 2021
Olaf Dössel, Tobias Schäffter, Gitta Kutyniok, Britta Rutert (Hrsg.)
Apps und Wearables für die Gesundheit (online und gedruckt)

8 / Dez 2021
Detlev Ganten, Max Löhning, Britta Rutert, Britta Siegmund
Gesundheitsregion Berlin-Brandenburg (online und gedruckt)

9 / Juni 2022
Jürgen Gerhards, Astrid Eichhorn, Julia Fischer, Ute Frevert und Christoph Markschies
Klimaschutz und akademische Dienstreisen. Empfehlungen für ein umweltschonendes Reiseverhalten in der Wissenschaft
(online und gedruckt)

10 / Juli 2022
Thomas Elsässer, Martin Grötschel, Matthias Scheffler, Joachim Hermann Ullrich, Friedhelm von Blanckenburg
Open Research Data in Naturwissenschaften und Mathematik: Empfehlungen der Mathematisch-Naturwissenschaftlichen Klasse der BBAW
(online)

In der Reihe „Denkanstöße“ werden Beiträge von Mitgliedern der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) zu aktuellen forschungspolitischen und wissenschaftlichen Themen veröffentlicht. Die namentlich gekennzeichneten Beiträge geben die Auffassung der Verfasserinnen und Verfasser wieder. Sie repräsentieren nicht notwendigerweise den Standpunkt der Akademie als Institution.

ISBN: 978-3-949455-18-6