FEDERATED MACHINE LEARNING ALS MITTEL ZUR ÜBERWINDUNG RECHTLICHER HÜRDEN DER FORSCHUNG MIT GESUNDHEITSDATEN

Jan Hospes / Walter Hötzendorfer / Phillipp Poindl / Christof Tschohl

Researcher, Research Institute AG & Co KG
Amundsenstraße 9, 1170 Wien, AT
jan.hospes@researchinstitute.at; https://researchinstitute.at

Senior Researcher, Research Institute AG & Co KG
Amundsenstraße 9, 1170 Wien, AT
walter.hoetzendorfer@researchinstitute.at; https://researchinstitute.at

Researcher, Research Institute AG & Co KG
Amundsenstraße 9, 1170 Wien, AT
phillipp.poindl @researchinstitute.at; https://researchinstitute.at

Wissenschaftlicher Leiter, Research Institute AG & Co KG
Amundsenstraße 9, 1170 Wien, AT
christof.tschohl@researchinstitute.at; https://researchinstitute.at

Schlagworte: ***Federated Machine Learning, KI, Gesundheitsdaten, Datenschutz, Privacy by Architecture, FeatureCloud***

Abstract: *Machine Learning in der medizinischen Forschung erfordert große Datenmengen und somit häufig die Einbeziehung von Daten aus mehreren medizinischen Einrichtungen. Federated Machine Learning ermöglicht dies, ohne dass die Daten die jeweilige Einrichtung verlassen, in der sie erhoben wurden. Dadurch steigert Federated Machine Learning nicht nur die Zugänglichkeit vorhandener Gesundheitsdaten für die Forschung, sondern eröffnet auch neue Möglichkeiten betreffend die Rechtsgrundlagen für diese Forschung, die im Beitrag analysiert werden.*

1. Einleitung

Grundvoraussetzung für Fortschritte im Bereich des Machine Learning (ML) und der darauf basierenden Forschung ist die Zugänglichkeit großer Datenmengen. Besonders im medizinischen Kontext existieren umfangreiche Datenbestände, deren Auswertung für neue Diagnosemethoden, das Verständnis von Krankheitsmechanismen, die Bewertung von Risikofaktoren etc. von entscheidender Bedeutung sein können. Häufig ist für ML jedoch methodenbedingt eine Anzahl relevanter Datensätze erforderlich, wie sie in einer einzelnen medizinischen Einrichtung nicht vorliegt bzw. nicht anfällt, beispielsweise in Form von elektronischen Gesundheitsakten (Electronic Health Records, EHR) oder Omics-Daten. Es ist daher geboten, Daten einzubeziehen, die über mehrere Einrichtungen verteilt sind. Dies ist eine erhebliche Herausforderung, da das Übermitteln der Daten in eine zentrale Datenbank häufig nicht möglich oder nicht tunlich ist. Die Gründe können sowohl objektive sein, wie rechtliche, organisatorische oder sonstige praktische Hürden oder auch subjektive, wie ein Mangel an Vertrauen. Soweit Letzteres nicht der Fall ist, besteht auf rechtlicher Ebene ein Lösungsweg aktuell in der Einholung eines Broad Consent. So hat die deutsche Medizininformatik-Initiative (MII)[1] unter Beachtung rechtlicher und ethischer Vorgaben einen Mustertext für die Einwilligung in die Sekundärnutzung

[1] https://www.medizininformatik-initiative.de/ (aufgerufen am 14.11.2023).

von Gesundheitsdaten erarbeitet und bereitgestellt, welche auch die Übermittlung von Daten einbezieht.[2] Dieser Mustertext wurde mit der deutschen Konferenz der unabhängigen Datenschutzbeauftragten des Bundes und der Länder abgestimmt, welche dazu ihr Einverständnis erklärt hat.[3]

Daneben zeigt die Europäischen Union mit Verordnung (EU) 2022/868[4] (Data Governance Act – DGA) durch die Einführung sog. Datenvermittlungsdienste Bestrebungen, Datenräume zu eröffnen, welche die gemeinsame Beforschung von Daten ermöglichen. Die Grundlage für die Einführung der Datenvermittlungsdienste findet sich in Art. 2 Z. 11 DGA, wonach es sich dabei um einen Dienst handelt, der durch technische, rechtliche oder andere Mittel Geschäftsbeziehungen zwischen einer unbestimmten Anzahl von betroffenen Personen oder Dateninhabern einerseits und Datennutzern andererseits herstellen soll. Im weiteren Verlauf der Norm erfolgt eine Negativabgrenzung, indem konkrete Dienste – etwa Clouddienste – ausgenommen werden. Ob der offenen Definition und der Klarstellung des ErwGr. 4 DGA, wonach die DSGVO[5] unberührt bleiben soll, ist aktuell noch nicht absehbar, ob die im DGA umschriebenen Dienste[6] die Datennutzung unter gleichzeitiger Wahrung hoher datenschutzrechtlicher Standards, wie von der EU-Kommission erhofft,[7] spürbar befördern werden.

2. Federated Machine Learning

Federated Machine Learning ermöglicht es, Machine Learning mit Daten durchzuführen, die über verschiedene Standorte verteilt sind, ohne die Daten in eine große gemeinsame Datenbank zu überführen. Stattdessen werden Daten nur lokal bei den Dateninhabern verarbeitet, im medizinischen Kontext somit in den Krankenhäusern oder sonstigen medizinischen Einrichtungen, im Folgenden als „Teilnehmer" bezeichnet, wo die Daten i.d.R. für Behandlungszwecke oder auch speziell für den jeweiligen Forschungszweck erhoben wurden. Dort wird ein sog. lokales Modell trainiert, wofür die Installation spezifischer Software und erforderlichenfalls auch Hardware notwendig ist. Anschließend werden die jeweiligen lokalen Modelle der einzelnen Teilnehmer an eine zentrale Stelle übermittelt, im Folgenden als Koordinator bezeichnet. Basierend auf den lokalen Modellen errechnet die Plattform des Koordinators ein globales Modell. Anders als bei obigen Lösungswegen liegt eine Neuerung auf technischer Ebene vor und es ist zu erforschen, welche rechtlichen Implikationen sich aus deren Einsatz ergeben.[8]

Grundlegendes Ziel des Federated Machine Learning ist es, das Trainieren von Modellen an Datenbeständen mehrerer Einrichtungen zu ermöglichen und hierbei gleichzeitig sicherzustellen, dass die zu diesen Zwecken verwendeten Rohdaten nicht die jeweilige Einrichtung verlassen.[9] Im Folgenden wird das im Horizon-Europe-Projekt FeatureCloud[10] entwickelte Federated-Machine-Learning-System beschrieben, welches sich als praxisorientierter Betrachtungsgegenstand eignet. Hinsichtlich der technischen Machbarkeit konnte im Forschungsprojekt FeatureCloud gezeigt werden, dass mittels Federated Machine Learning vergleichbare

---

2 MII Arbeitsgruppe Consent, Mustertext Patienteneinwilligung (Stand 16.04.2020), https://www.medizininformatik-initiative.de/sites/default/files/2020-04/MII_AG-Consent_Einheitlicher-Mustertext_v1.6d.pdf (aufgerufen am 14.11.2023).

3 MII, Medizininformatik-Initiative erhält grünes Licht für bundesweite Patienteneinwilligung, https://www.medizininformatik-initiative.de/de/medizininformatik-initiative-erhaelt-gruenes-licht-fuer-bundesweite-patienteneinwilligung (aufgerufen am 14.11.2023).

4 Verordnung (EU) 2022/868 des Europäischen Parlaments und des Rates vom 30. Mai 2022 über europäische Daten-Governance und zur Änderung der Verordnung (EU) 2018/1724, ABl L 2022/152.

5 Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG, ABl L 2016/119.

6 Bitkom, Das sind die neuen Pflichten für Datenvermittlungsdienste durch den Data Governance Act (Stand 2023 Version 2), https://www.bitkom.org/sites/main/files/2023-05/BitkomDGADVDOrientierungshilfe2023.pdf (aufgerufen am 14.11.2023).

7 https://digital-strategy.ec.europa.eu/de/policies/data-governance-act-explained (aufgerufen am 14.11.2023).

8 Diese sowie die Forschung zu diesem Beitrag erfolgte im Forschungsprojekt FEATURECLOUD (https://featurecloud.eu/), finanziert im Förderprogramm HORIZON 2020 der Europäischen Union unter der Finanzhilfevereinbarung Nr. 826078.

9 Yang/Liu/Chen/Tong, Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), volume 10, issue 2, 2019, S. 1–19. DOI: 10.1145/3298981

10 https://featurecloud.eu (aufgerufen am 14.11.2023).

Ergebnisse erzielt werden können, wie mittels State-of-the-Art-Machine-Learning-Methoden auf zentral zusammengeführten Datenbeständen.[11] Mit dem FeatureCloud App Store[12] wurde eine Plattform für die Entwicklung und Veröffentlichung datenschutzfreundlicher Federated-Machine-Learning-Algorithme geschaffen.

Der Machine-Learning-Prozess läuft hier, vereinfacht dargestellt, wie folgt ab: Zunächst wählt ein Koordinator die für ein Forschungsprojekt anzuwendenden Machine-Learning-Algorithmen aus und determiniert damit wesentlich die Forschungsziele. Einrichtungen, welche dem Forschungsvorhaben Daten zur Verfügung stellen möchten, übermittelt der Koordinator einen Einladungstoken. Daraufhin wählen die Einrichtungen aus ihren Datenbeständen projektrelevante Daten aus und errechnen unter Verwendung der durch den Koordinator vorgegebenen Trainingsalgorithmen auf der unter ihrer vollständigen Kontrolle stehenden Infrastruktur lokale Modelle. Nachdem ein Teilnehmer einen Lernvorgang abgeschlossen hat, sendet er das daraus resultierende lokale Modell an den Koordinator, welcher aus der Gesamtheit aller so empfangenen Modelle ein gemeinschaftliches „globales" Modell errechnet.[13]

3. Personenbezug im Federated-Machine-Learning-System

Wie oben dargestellt, liegen die Rohdaten ausschließlich den Teilnehmern vor. Diese können personenbezogen in den Lernprozess einfließen oder bereits auf lokaler Ebene anonymisiert oder pseudonymisiert werden, wobei für die weitere Beurteilung vom Vorliegen eines Personenbezugs ausgegangen wird, sodass überhaupt die Möglichkeit besteht, dass lokale Modelle einen Personenbezug aufweisen können.

Gemäß ErwGr 26 DSGVO liegt ein Personenbezug vor, wenn unter Berücksichtigung aller Mittel, die nach allgemeinem Ermessen wahrscheinlich genutzt werden können, die Zuordnung der Daten zu einer natürlichen Person möglich ist. Die relative Auslegung des Personenbezugs, welcher der EuGH folgt, stellt darauf ab, ob ein Verantwortlicher aufgrund seiner individuellen Kenntnisse, Mittel und Möglichkeiten den Bezug zur betroffenen Person herstellen kann.[14] Daten sind für den Verantwortlichen, der in der Lage ist, diese mit einer bestimmten Person in Verbindung zu bringen, weiterhin personenbezogen. Hingegen sind dieselben Daten für Dritte, die nicht im Besitz der Zuordnungsregel und/oder anderer Mittel sind, anonym.

Modelle können personenbezogene Details über die Trainingsdaten speichern, die nichts mit der beabsichtigten Aufgabe zu tun haben.[15] Böswillige Algorithmen für Machine Learning können Modelle erstellen, die eine erhebliche Menge an Informationen über ihre Trainingsdatensätze preisgeben, selbst wenn der Angreifer nur Blackbox-Zugriff auf das Modell hat.[16]

Im Kontext des Federated Machine Learning bedeutet dies, dass eine praktische Bewertung des Personenbezugs durchgeführt werden muss. Von allen Angriffsvektoren auf die Anonymität der Daten sind all jene, die in der Praxis vernünftigerweise von einem tatsächlichen Angreifer verwendet werden können, auf der Grundlage objektiver Faktoren wie den Kosten und der benötigten Zeit, den erforderlichen Fähigkeiten, dem

---

[11] Nasirigerdeh/Torkzadehmahani/Matschinske/Frisch/List/Späth/Weiss/Völker/Pitkänen/Heider/Wenke/Kaissis/Rueckert/Kacprowski/Baumbach, sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies, Genome Biology, volume 32, issue 1, 2022. DOI: 10.1186/s13059-021-02562-1; Zolotareva/Nasirigerdeh/Matschinsk/Torkzadehmahani/Bakhtiari/Frisch/Späth/Blumenthal/Abbasinejad/Tieri/Kaissis/Rückert/Wenke/List/Baumbach/Flimma, a federated and privacy-aware tool for differential gene expression analysis, Genome Biology, volume 22, issue 1,2021. DOI: 10.1186/s13059-021-02553-2.

[12] https://featurecloud.ai (aufgerufen am 14.11.2023).

[13] Matschinske/Späth/Nasirigerdeh/Fejér, Deliverable D7.2 "App store ready and extendible by developers", https://featurecloud.eu/wp-content/uploads/2021/01/D7.2_App_store_ready_and_extendible_by_developers.pdf (aufgerufen am 14.11.2023).

[14] EuGH 19.10.2016, C-582/14.

[15] Carlini/Liu/Erlingsson/Kos/D. Song, The secret sharer: Evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Conference on Security Symposium. USENIX Association, Santa Clara 2019, S. 16.

[16] Song/Ristenpart/Shmatikov, Machine Learning Models that Remember Too Much. In: CCS ,17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, New York 2017, S. 587 ff.

potenziellen Gewinn und der verfügbaren Technologie, aber auch möglichen technologischen Entwicklungen in der Zukunft, zu bewerten. Hinreichend unwahrscheinliche Angriffe, also solche, bei denen nicht vorstellbar ist, dass sie in der Praxis in dem gegebenen Kontext stattfinden, etwa weil der Angreifer den erforderlichen Aufwand als unverhältnismäßig hoch einstufen wird, fließen nicht in die Bewertung des Personenbezugs ein.

Die Herstellung des Personenbezugs ist erschwert, wenn das Modell mit nicht exakt reproduzierbaren Daten trainiert wurde. Daher sollte bei der Analyse auch zwischen exakt reproduzierbaren und nicht exakt reproduzierbaren Daten unterscheiden werden. Nicht exakt reproduzierbare Daten sind insbesondere solche, deren Erhebung einer gewissen Messungenauigkeit unterliegt, und/oder deren zugrundeliegende Werte, z.B. Blutwerte, im Laufe der Zeit veränderlich sind, sodass eine erneute medizinische Untersuchung zu einem späteren Zeitpunkt nicht zu exakt denselben Daten führt. Gegenbeispiele sind etwa die Blutgruppe und insbesondere genetische Daten, die (theoretisch) bei einer neuerlichen Erhebung exakt reproduzierbar sind.

Als Zwischenergebnis ist festzuhalten, dass es erreichbar ist, dass lokale Modelle nicht personenbezogen sind und der Koordinator damit ausschließlich Daten verarbeitet, die keinen Personenbezug aufweisen. Ob und welche Maßnahmen (zB Differential Privacy, Secure Multiparty Computation)[17] zu ergreifen sind, damit dieser Zustand hergestellt werden kann, ist im Einzelfall zu entscheiden.

4. Komplexität der Rollenverteilung

Um eine auf „faktischen Elementen oder Umständen“ basierende Zuordnung der zentralen Rolle des Verantwortlichen erreichen zu können, sollte im Einklang mit der bereits von der Art-29-Datenschutzgruppe entwickelten Ermittlungsmethodik zunächst die jeweilige Verarbeitungstätigkeit isoliert betrachtet und ermittelt werden, warum diese Verarbeitung überhaupt durchgeführt wird. Sofern zwei oder mehr Verantwortliche gemeinsam die Zwecke und Mittel der Verarbeitung festlegen, führt dies zur sogenannten “pluralistische[n] Kontrolle”[18] über die jeweilige Datenverarbeitungstätigkeit, womit die gemeinsame Verantwortlichkeit nach Maßgabe von Art. 26 DSGVO begründet ist.

Die Kriterien für das Vorliegen gemeinsamer Verantwortlichkeit wurden durch die Rechtsprechung häufig erweitert.[19] Bei weiter Interpretation der Rechtssache Fashion ID könnte der Schluss gezogen werden, dass jeder Akteur, der die Verarbeitung personenbezogener Daten ermöglicht, als für die Verarbeitung Mitverantwortlicher gilt.[20] Nach dem EDSB ist das übergreifende Kriterium für das Vorliegen einer gemeinsamen Verantwortlichkeit die gemeinsame Beteiligung von zwei oder mehr Stellen an der Festlegung der Zwecke und Mittel einer Verarbeitung. Eine gemeinsame Festlegung bedeutet (unter anderem) eine gemeinsame Entscheidung, was bedeutet, dass die Akteure gemeinsam entscheiden und eine gemeinsame Absicht haben.[21] Ob es eine gemeinsame Kontrolle zwischen Koordinator und Teilnehmern gibt, hängt allgemein weitgehend von der Gestaltung der Vorprojektphase ab. Durch die Auswahl der Lernalgorithmen und die Zusammenstellung der Arbeitsabläufe prädeterminiert der Koordinator standardmäßig die Zwecke der Verarbeitung.

Die Zusammenarbeit des Teilnehmers mit dem Koordinator, die sich aus der Auswahl der konkreten Daten durch den Teilnehmer ergibt, führt so jedenfalls zu einer gemeinsamen Verantwortlichkeit. Da ein Teilnehmer jedoch üblicherweise weder Zweck noch Verarbeitungsmittel gemeinsam mit anderen Teilnehmern festlegt oder Einfluss auf deren Zusammensetzung hat, ist in der Regel keine gemeinsame Verantwortlichkeit

---

17 Winter/Steinebach/Heereman/Steiner/Battis/Halvani/Yannikos/Schüssler, Privacy und Big Data, Fraunhofer-Institut für Sichere Informationstechnologie SIT, Darmstadt, 2020, S. 78, 91 ff.

18 Artikel-29-Datenschutzgruppe, Stellungnahme 1/2010 zu den Begriffen „für die Verarbeitung Verantwortlicher“ und „Auftragsverarbeiter“ (WP 169), S. 38 ff; Hödl In: Knyrim (Hrsg.), DatKomm, 2018, Art. 4 Rz. 80.

19 Millard/Kuner/Cate/Lynskey/Loideain/Svantesson, At This Rate, Everyone Will Be a [Joint] Controller of Personal Data!, International Data Privacy Law, volume 9, issue 4, 2019, S. 217–219. DOI: 10.1093/idpl/ipz027.

20 Bobek, ECLI:EU:C:2018:1039, Rz. 74.

21 EDPB, Guidelines 07/2020 on the concepts of controller and processor in the GDPR (September 2020 Version 1.0), S. 17 ff.

der Teilnehmer untereinander anzunehmen. So implizieren auch die spanische Datenschutzbehörde und der Europäische Datenschutzbeauftragte in einem gemeinsam formulierten Beitrag, dass die Teilnehmer an einem föderierten System für Machine Learning als eigenständige für die Verarbeitung Verantwortliche zu qualifizieren sind.[22] Im Falle einer Organisation, welche eine Verarbeitungstätigkeit organisiert und koordiniert, und mehreren an sie übermittelnden Stellen geht auch der EuGH nicht von einer gemeinsamen Verantwortlichkeit zwischen den übermittelnden Stellen, sondern von gemeinsamer Verantwortlichkeit der Organisation jeweils mit den einzelnen übermittelnden Stellen aus.[23] Zwar liegt ein gewisses Maß an „Interessensgleichrichtung"[24] der Teilnehmer vor, allerdings erschöpft sich diese vor allem in dem Ziel des allseitigen Einbringens der jeweiligen lokalen Daten.

Somit hat der Koordinator eine Vereinbarung gemäß Art. 26 Abs. 1 und 2 DSGVO mit den Teilnehmern zu treffen. Darin ist klar festzulegen, dass eine gemeinsame Verantwortlichkeit vorliegt, wie jeder der Verantwortlichen an der Entscheidung über die Zwecke und Mittel der gemeinsamen Verarbeitung mitwirkt und wer von den Verantwortlichen welche Verpflichtungen nach der DSGVO zu erfüllen hat.[25]

Durch die maßgebliche Determinierungsfunktion des Koordinators liegt die gemeinsame Verantwortlichkeit in der Regel somit jeweils zwischen Koordinator und den einzelnen Teilnehmern, nicht jedoch zwischen den Teilnehmern vor. Somit erfordert die Hinzunahme eines Teilnehmers lediglich den Abschluss einer einzigen Vereinbarung über die gemeinsame Verantwortlichkeit (mit dem Koordinator), nicht jedoch mit den anderen Teilnehmern, was ansonsten mit steigender Zahl an Teilnehmern zu einer Explosion des dafür erforderlichen Aufwands führen würde.

5. Rechtsgrundlagen der Datenverarbeitung

Jede Form der Nutzung personenbezogener Daten und damit jedenfalls auch die datengetriebene Gesundheitsforschung, Entwicklung sowie Validierung von Machine Learning-basierten Verfahren, erfordert eine rechtliche Grundlage in Form eines Erlaubnistatbestands der DSGVO.

Die Rechtsgrundlage der Einwilligung ist hierbei als normative Ausprägung des Grundsatzes der informationellen Selbstbestimmung[26] von zentraler Bedeutung. So vertritt etwa der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (BfDI) die Auffassung, dass es dem Schutz der sensiblen und besonders zu schützenden Gesundheitsdaten am ehesten gerecht wird, wenn sogar nationale gesetzliche Grundlagen für die Nutzung von Gesundheitsdaten zu Forschungszwecken, welche an sich bereits als eine Rechtsgrundlage für die Datenverarbeitung dienen könnten, eine Zustimmung der Betroffenen als Zulässigkeitsvoraussetzung vorsehen.[27] Auch wenn in Art. 8 Abs. 2 GRC die Einwilligung ausdrücklich angeführt wird und im Übrigen „nur" sonstige gesetzlich geregelte legitime Grundlagen als Voraussetzung für eine Verarbeitung genannt werden, ist zu betonen, dass zwischen den Erlaubnistatbeständen in Art. 6 Abs. 1 DSGVO kein Rangverhältnis besteht.[28]

---

22 AEPD, EDPS, Joint Paper – 10 Misunderstandings about Machine Learning, https://edps.europa.eu/data-protection/our-work/publications/papers/2022-09-20-aepd-edps-joint-paper-10-misunderstandings-about-machine-learning (aufgerufen am 14.11.2023).

23 EuGH 10.07.2018, C-25/17.

24 EuGH 10.07.2018, C-25/17, Rz. 68 ff.

25 Veil In: Gierschmann/Schlender/Stentzel/Veil (Hrsg.), DS-GVO, 2017, Art. 26 Rz. 64.

26 Buchner/Petri In: Kühling and Buchner (Hrsg.), DS-GVO/BDSG, 2018, Art. 6 Rz. 17.

27 https://www.bfdi.bund.de/DE/Buerger/Inhalte/GesundheitSoziales/Allgemein/MedizinischeForschung.html (aufgerufen am 14.11.2023).

28 Kastelitz/Hötzendorfer/Tschohl In: Knyrim (Hrsg.), DatKomm, 2018, Art. 6 Rz. 14.

5.1. Verarbeitung durch den Teilnehmer

Der Teilnehmer verarbeitet im Zuge des Trainings des lokalen Modells personenbezogene Daten und muss daher in diesem Zusammenhang eine Rechtsgrundlage festlegen. Dem Teilnehmer steht prinzipiell die gesamte Palette der Rechtsgrundlagen der DSGVO zur Verfügung. In der Praxis wird die Rechtsgrundlage meist entweder in der Weiterverarbeitung für kompatible Zwecke nach Art. 6 Abs. 4 DSGVO (Sekundärnutzung) in Bezug auf Daten, die im Rahmen eines Behandlungsvertrags erhoben wurden, oder in der Einholung einer ausdrücklichen Einwilligung nach Art. 9 Abs. 2 lit. a DSGVO oder im auf Basis von Art. 9 Abs. 2 lit. j DSGVO durch die nationale Rechtsordnung näher ausgestalteten Forschungsprivileg zu finden sein.

In der Praxis wird für medizinische Forschungszwecke häufig die Einwilligung als Rechtsgrundlage herangezogen. Unter datenschutzrechtlichen Laien wird diese erfahrungsgemäß vielfach sogar als die einzige denkbare Rechtsgrundlage angesehen. Sie wird aus ethischer Sicht dem Ideal der Selbstbestimmung der betroffenen Person wohl auch am besten gerecht. In der Praxis wird jedoch in vielen Fällen zu hinterfragen sein, inwieweit die erforderliche Einsicht in die Art und die Risiken der Datenverarbeitung sowie die erforderliche Freiwilligkeit der Einwilligung tatsächlich gegeben sind, und sei es „nur" deswegen, weil viele betroffene Personen sich dafür nicht ausreichend Zeit nehmen oder nehmen können. Somit erscheint es auch aus ethischer Sicht angemessen, die erforderliche Abwägung dem Gesetzgeber zu überlassen und die Verarbeitung auf einen Erlaubnistatbestand zu stützen, den der Gesetzgeber unabhängig vom aktiven Zutun der betroffenen Person eingeräumt und entsprechend ausgestaltet hat. Die Informationspflichten nach Art. 13 f. DSGVO bestehen ohnehin auch in diesen Fällen, worauf an dieser Stelle ausdrücklich hingewiesen sei. Im Folgenden soll auf die Weiterverarbeitung für kompatible Zwecke nach Art. 6 Abs. 4 DSGVO näher eingegangen werden, die in besonderer Weise mit Federated Machine Learning kompatibel erscheint, weil Federated Machine Learning nicht die Verarbeitung der Daten durch einen anderen Verantwortlichen als ursprünglichen Dateninhaber erfordert, wie im nächsten Abschnitt noch gezeigt werden wird.

Der Grundsatz der Zweckbindung in Art. 5 Abs. 1 lit. b DSGVO ist ein Kernbestandteil des europäischen Datenschutzrechts und auch primärrechtlich in Art. 8 Abs. 2 GRCh verankert. Zweckbindung bedeutet, dass personenbezogene Daten nur für (vorab) festgelegte, eindeutige und rechtmäßige Zwecke erhoben und nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden dürfen.[29] Allerdings wird bereits im zweiten Halbsatz von Art. 5 Abs. 1 lit. b DSGVO die Möglichkeit einer (zweckändernden) Weiterverarbeitung zu kompatiblen Zwecken normiert.[30] Art. 6 Abs. 4 DSGVO regelt ausdrücklich die Weiterverarbeitung personenbezogener Daten im Sinne einer „Sekundärnutzung". Insofern stellt die Vorschrift des Art. 6 Abs. 4 DSGVO eine normative Durchbrechung des strikten Zweckbindungsgrundsatzes dar.

Ob eine Weiterverarbeitung zu kompatiblen Zwecken vorliegt und somit zulässig ist, ist grundsätzlich anhand der in Art. 6 Abs. 4 DSGVO angeführten Kriterien zu prüfen. Für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke stellt Art. 5 Abs. 1 lit. b, dritter Halbsatz DSGVO unter Bezugnahme auf Art. 89 Abs. 1 DSGVO die (gesetzliche) Fiktion (praesumptio iuris ac de iure) auf,[31] wonach die Weiterverarbeitung (auch von „sensiblen Daten") für diese Zwecke nicht als unvereinbar mit den ursprünglichen Zwecken angesehen wird.[32] Gestützt auf den klaren Wortlaut des Art. 5 Abs. 1 lit. b DSGVO wird in Teilen der Literatur argumentiert, dass bei Vorliegen eines dieser Zwecke eine Vereinbarkeitsprüfung gar nicht durchgeführt werden muss.[33] *Buchner* wiederum

[29] Hötzendorfer/Tschohl/Kastelitz In: Knyrim (Hrsg.), DatKomm, 2018, Art. 5 Rz. 20.

[30] Kastelitz/Hötzendorfer/Tschohl In: Knyrim (Hrsg.), DatKomm, 2018, Art. 6 Rz. 58.

[31] Kotschy, Die Zulässigkeitsvoraussetzungen für Forschungsdatenverarbeitungen nach dem FOG – eine kritische Analyse. In: Jahnel (Hrsg.), Jahrbuch Datenschutzrecht 2020, S. 287.

[32] Gabauer, Die Verarbeitung personenbezogener Daten zu wissenschaftlichen Forschungszwecken, NWV im Verlag Österreich, 2019, S. 53.

[33] Kastelitz/Hötzendorfer/Tschohl In: Knyrim (Hrsg.), DatKomm, 2018, Art. 6 Rz. 64; Reimer In: Sydow (Hrsg.), DS-GVO, 2018 Art. 5 Rz. 27.

hebt hervor, dass diese „Öffnung" als Ausnahme vom Zweckbindungsgrundsatz von vornherein eng zu verstehen sei, weshalb der pauschale Verweis auf wissenschaftliche, historische oder statistische Zwecke nicht ausreichen könne.[34] *Roßnagel* argumentiert, dass sich die – die Zweckänderung erleichternde – rechtliche Fiktion nicht aus der generellen Höherwertigkeit der vier Zwecke ergibt, sondern vielmehr „führen diese spezifischen Verwendungszwecke dazu, dass sich die Datenverarbeitung typischer Weise nicht auf die Person bezieht, deren Daten verarbeitet werden. Personenbezogene Daten sind der Ausgangspunkt der Verarbeitung, aber nicht das Ergebnis."[35] Für statistische Zwecke wird dies in ErwGr. 162 der DSGVO ausdrücklich festgestellt. Die Fiktion gelte daher nicht für alle Verarbeitungstätigkeiten, die wissenschaftliche, historische oder statistische Methoden verwenden, sondern nur für solche, die auf nicht personenbezogene Ergebnisse abzielen.[36] Aufgrund der doppelten Verneinung des Art. 5 Abs. 1 lit. b DSGVO „nicht als unvereinbar", spricht bei Vorliegen von Forschungszwecken nach *Roßnagel* eine Vermutung für die Kompatibilität der Zwecke, jedoch hat auch in diesem Fall eine Einzelfallprüfung der Vereinbarkeit mit dem Erhebungszweck zu erfolgen.[37] Mit Blick auf die ausgeführten Hintergründe der rechtlichen Fiktion kann dies aber nur bedeuten, dass sich die Vereinbarkeitsprüfung darin erschöpfen muss, zu prüfen, ob der Weiterverarbeitungszweck tatsächlich auf nicht personenbezogene Ergebnisse abzielt. Dies ist bei Federated Machine Learning stets der Fall, sofern die lokalen Modelle – welche die Ergebnisse der Verarbeitung sind – keinen Personenbezug aufweisen, was oben in Abschnitt 3 näher erläutert wurde.

Zu beachten ist auch, dass bei einer solchen zulässigen Weiterverarbeitung die in Art. 89 Abs. 1 DSGVO genannten Garantien zum Schutz der Grundrechte und Grundfreiheiten der betroffenen Personen berücksichtigt werden müssen. Zudem muss der für die Verarbeitung Verantwortliche die betroffene Person gemäß Art. 13 Abs. 3 und Art. 14 Abs. 4 über die Änderung des Zwecks informieren. Dies gilt auch für Zweckänderungen, die mit dem Zweck der Erhebung vereinbar sind.[38]

In der Literatur ist umstritten, ob die Verarbeitung zu kompatiblen Zwecken einer gesonderten Rechtsgrundlage bedarf oder nicht.[39] Erwägungsgrund 50 Satz 2 der DSGVO bringt jedoch klar zum Ausdruck, dass im Fall der Verarbeitung zu kompatiblen Zwecken „keine andere gesonderte Rechtsgrundlage erforderlich [ist] als diejenige für die Erhebung der personenbezogenen Daten". Diese in der deutschen Fassung u.E. etwas missglückte Formulierung lautet in der englischen Fassung wie folgt: „In such a case, no legal basis separate from that which allowed the collection of the personal data is required."

Folgt man diesen u.E. naheliegenden Auslegungen, dann lässt sich das Ergebnis wie folgt zusammenfassen: Erfolgt die Weiterverarbeitung von bestehenden Daten für wissenschaftliche Forschungszwecke durch denselben für die Verarbeitung Verantwortlichen ausschließlich zur Erzielung von Ergebnissen, die keine personenbezogenen Daten enthalten, dann ist sie gemäß Art. 6 Abs. 4 DSGVO rechtmäßig.

5.2. Verarbeitung durch den Koordinator

Der Koordinator erhält im Rahmen des Federated Machine Learning von den Teilnehmern das jeweils lokal bei jedem einzelnen Teilnehmer trainierte Modell. Unter der oben diskutierten Voraussetzung, dass dieses keine personenbezogenen Daten enthält, verarbeitet der Koordinator keine personenbezogenen Daten. Die perso-

[34] Buchner, Grundsätze und Rechtmäßigkeit der Datenverarbeitung unter der DS-GVO Datenschutz und Datensicherheit, Datenschutz und Datensicherheit – DuD, 2016, Heft 4, S. 157.

[35] Rossnagel In: Simitis/Hornung/Spiecker (Hrsg.), Datenschutzrecht, 2019, Art. 5 Abs. 1 Rz. 104.

[36] Rossnagel In: Simitis/Hornung/Spiecker (Hrsg.), Datenschutzrecht, 2019, Art. 5 Abs. 1 Rz. 104; Siehe auch: Artikel 29 Datenschutzgruppe, Opinion 03/2013 on purpose limitation (WP 203), S. 28.

[37] Rossnagel In: Simitis/Hornung/Spiecker (Hrsg.), Datenschutzrecht, 2019, Art. 6 Abs. 4 Rz. 41.

[38] Rossnagel In: Simitis/Hornung/Spiecker (Hrsg.), Datenschutzrecht, 2019, Art. 6 Abs. 4 Rz. 47.

[39] Dafür spricht: Herbst in Kühling/Buchner (Hrsg.), DS-GVO BDSG3 Art. 5 Rn 54; andere Ansicht, gegen die Notwendigkeit einer gesonderte Rechtsgrundlage: Rossnagel In: Simitis/Hornung/Spiecker (Hrsg.), Datenschutzrecht, 2019, Art. 5 Abs. 1 Rz. 98 f; Kastelitz/Hötzendorfer/Tschohl In: Knyrim (Hrsg.), DatKomm, 2018, Art. 6 Rz. 62.

nenbezogenen Daten werden lediglich lokal bei und von den einzelnen Teilnehmern verarbeitet, mit denen er, wie oben erläutert, jeweils gemeinsam für die Verarbeitung ihrer jeweiligen lokalen Daten verantwortlich ist. Eine gemeinsame Verantwortlichkeit stellt weder eine Rechtsgrundlage für die Verarbeitung durch mehrere Verantwortliche dar, noch ist eine Rechtsgrundlage dafür erforderlich, dass mehrere Verantwortliche eine gemeinsame Verantwortlichkeit eingehen. Soweit ein bestimmter Verantwortlicher im Rahmen der gemeinsamen Verantwortlichkeit personenbezogene Daten verarbeitet, benötigt dieser bestimmte Verantwortliche eine eigene Rechtsgrundlage für diese Verarbeitung.[40] Dies wird auch durch den Wortlaut der Art. 6 und 9 DSGVO gestützt, die eindeutig festlegen, dass für die „Verarbeitung" personenbezogener Daten eine Rechtsgrundlage erforderlich ist. Der Verarbeitungsbegriff wird durch eine taxative Aufzählung von Prozessen in Art. 4 Abs. 2 DSGVO präzisiert. Umfasst sind neben dem bloßen Innehaben der Daten nur Handlungen mit den Daten, d.h. Vorgänge mit unmittelbaren Auswirkungen auf die Daten und keine Handlungen, die sich nur indirekt auf die Daten auswirken können, wie z.B. die Festlegung der Zwecke der tatsächlich von einer anderen Stelle durchgeführten Datenverarbeitung. Zusammenfassend lässt sich sagen, dass eine genaue Analyse zu dem Schluss führt, dass die DSGVO das Erfordernis einer Rechtsgrundlage an die tatsächliche Verarbeitung personenbezogener Daten und nicht an die Rolle des (gemeinsam) für die Verarbeitung Verantwortlichen knüpft.

Ein Koordinator, der mit einem Teilnehmer eine gemeinsame Verantwortlichkeit eingeht, aber selbst keine personenbezogenen Daten verarbeitet, wie oben beschrieben, muss daher für die Durchführung von Federated Machine Learning nicht über eine eigene Rechtsgrundlage verfügen. Nur jeder einzelne Teilnehmer muss für die tatsächliche Verarbeitung der jeweils lokalen Daten, wie oben beschrieben, über eine Rechtsgrundlage gemäß der DSGVO verfügen, unabhängig vom Vorhandensein eines zentralen Koordinators.

6. Fazit

Federated Machine Learning erscheint geeignet, wesentliche rechtliche und praktische Hürden bei kollaborativer medizinischer Forschung zu überwinden, insbesondere weil dabei die beforschten Daten die jeweilige medizinische Einrichtung, in der sie erhoben wurden, nicht verlassen. Dies trifft allerdings nur zu, wenn auch die von den Teilnehmern an den Koordinator übermittelten lokal trainierten Modelle keine personenbezogenen Daten enthalten, sodass der Koordinator selbst zu keinem Zeitpunkt personenbezogene Daten verarbeitet. Nicht zuletzt durch die Anwendung von Privacy Enhancing Technologies wie Secure Multi-Party Computation (SMPC) und/oder Differential Privacy (DP) ist dies auch praktisch umsetzbar. Unter dieser Voraussetzung kann Federated Machine Learning – wie gezeigt wurde – ausschließlich auf Basis von Rechtsgrundlagen, die für die Verarbeitung der jeweils lokal vorhandenen Daten in den einzelnen beteiligten medizinischen Einrichtungen vorliegen, durchgeführt werden, ohne dass der Koordinator einer eigenen Rechtsgrundlage bedarf. Dies deshalb, weil die DSGVO das Erfordernis einer Rechtsgrundlage an die tatsächliche Verarbeitung personenbezogener Daten und nicht an die Rolle des (gemeinsam) für die Verarbeitung Verantwortlichen knüpft.

[40] Datenschutzkonferenz, Kurzpapier Nr. 16 Gemeinsam für die Verarbeitung Verantwortliche, Art. 26 DS-GVO, https://www.datenschutzkonferenz-online.de/media/kp/dsk_kpnr_16.pdf (aufgerufen am 14.11.2023), S. 1.