Rainer Mühlhoff

Die Illusion der Anonymität: Big Data im Gesundheitssystem

Im Streit um die Corona-App traten zivilgesellschaftliche Akteure erbittert für mehr Datenschutz ein – und ihr Erfolg hat deutlich gezeigt: Gesundheits- und Datenschutz stellen keinen Widerspruch dar, sondern können Hand in Hand gehen.[1]

Dass sich diese Erkenntnis ausgerechnet in der Pandemie durchsetzte, ist eine große Chance für den Datenschutz im Gesundheitsbereich. Denn in den meisten Debatten wird Datenschutz noch immer als Hindernis für geradezu revolutionäre Innovationen dargestellt. So versprechen sich Politiker*innen von einer umfassenden elektronischen Verarbeitung der Patient*innendaten eine deutlich verbesserte Gesundheitsversorgung. In der medizinischen Forschung soll die Auswertung großer Datensätze (Big Data) mittels „Data Mining" und Künstlicher Intelligenz sogar gänzlich neue Diagnose- und Therapiemöglichkeiten erschließen. Hier wie dort werden Datenschutzbedenken zumeist relativiert oder – nicht zuletzt im internationalen Kontext – als spezifisch deutsche Befindlichkeit abgetan.

Tatsächlich aber sollte der Datenschutz gerade in der Gesundheitsversorgung nicht als eine Befindlichkeit, sondern als essenzieller Schutz unserer Grundrechte geachtet werden. Um diesen Schutz zu gewähren, wird häufig von Dienstanbietern eine Pseudo- oder Anonymisierung sensibler Daten zugesagt. Allerdings können selbst diese Daten noch dazu verwendet werden, Informationen über unsere Gesundheit abzuleiten – vor allem dann, wenn pseudonymisierte Informationen von vielen Millionen Patient*innen vorliegen: Durch Mustererkennung und Künstliche Intelligenz können aus Massendaten sensible Daten einzelner Personen mittels Schätzung abgeleitet werden, zum Beispiel ob wir alkoholabhängig sind oder zu Depressionen neigen. Insbesondere mit Blick auf die Gesundheitsversorgung benötigen wir daher ein erweitertes Verständnis von Datenschutz, das auch die Verwendung anonymisierter und pseudonymisierter Daten strenger reguliert.[2]

Lecks und Bugs

Daten, die unsere Gesundheit betreffen, sind oft schützenswerter als andere Daten, weil sich viele medizinische Befunde ein Leben lang nicht ändern. Gelangen solche Informationen in die falschen Hände, können sie demnach über Jahrzehnte zum Nachteil der Betroffenen verwendet werden.

Entsprechend wertvoll sind Patientendaten auf dem (Schwarz-)Markt der Datenhändler und Analyseunternehmen wie Cambridge Analytica.[3] Allein in den vergangenen Monaten gab es mehrere Datenlecks im Gesundheits-

1 Vgl. Daniel Leisegang, Corona und die Grundrechte: Einsicht in die Notwendigkeit, in: „Blätter", 5/2020, S. 25-28.

2 Anonymisierte Daten sind solche, die keinerlei Personenbezug mehr aufweisen, etwa weil Gruppen oder Cluster statistisch zusammengefasst wurden. Von Pseudonymisierung spricht man hingegen, wenn Chiffren oder Nummern statt eines Identifikationsmerkmals verwendet werden.

3 Vgl. The Cambridge Analytica Files, www.theguardian.com.

system, die solchen Akteuren überaus gelegen kommen. So wurde im Februar bekannt, dass Millionen Krankenakten deutscher Arztpraxen aufgrund schlecht geschützter Server im Internet zugänglich waren. Etwa zur gleichen Zeit berichteten Medien über Einsatzdaten des Deutschen Roten Kreuzes in Brandenburg, die gänzlich ungesichert auf einem Internetserver abgelegt worden waren. Und bereits 2018 hatten Hacker der norwegischen Gesundheitsbehörde Patient*innendaten von mehr als der Hälfte der norwegischen Bevölkerung entwendet.

Diese Fälle zeigen einmal mehr, dass elektronische Datenspeicherung stets mit inhärenten Risiken verbunden ist, die eine ständige Überprüfung der Systeme erfordern. Auch das pauschale Versprechen einer „Verschlüsselung" der Daten reicht da nicht aus: Die angewandten Verfahren veralten mit der Zeit und müssen durch neue, sicherere ersetzt werden. Auch Programmierfehler – sogenannte Bugs – werden meist zufällig entdeckt und müssen dann umgehend und flächendeckend behoben werden, um größeren Schaden abzuwenden. Und schließlich führt nicht selten auch die Fehlkonfiguration sicherheitsrelevanter Software dazu, dass Daten frei zugänglich sind.

Risiko elektronische Patientenakte

Grundsätzlich ist der drohende Schaden umso größer, je mehr Datensätze an ein und demselben Ort gespeichert werden. Ein solch hohes Risiko liegt etwa auch bei der „elektronischen Patientenakte" (ePA) vor. Nach den Plänen des Gesundheitsministeriums soll sie im kommenden Jahr bundesweit als Erweiterung der „elektronischen Gesundheitskarte" (eGK) eingeführt werden. Während die eGK nur die Stammdaten der Patient*innen, gegebenenfalls ihre Notfalldaten (Kontakte, Blutgruppe, Allergien, Organspendebereitschaft etc.) sowie einen Medikationsplan auf der Karte selbst – also dezentral – speichert, sollen die Daten der ePA auf einem zentralen Server abgelegt werden. Die ePA bricht also mit dem Prinzip der dezentralen Speicherung und physischen Überlassung des Speichermediums an die Dateneigentümer*innen und ermöglicht dadurch potentiell die gesammelte Entwendung vieler Datensätze durch einen einzigen Einbruch auf dem Server.

Ein zentraler Zugriff auf die Daten aller gesetzlich Versicherten in Deutschland soll künftig auch für Forschungszwecke möglich werden. Das vom Deutschen Bundestag Ende 2019 beschlossene „Digitale-Versorgung-Gesetz" schafft den Rahmen für die Zusammenführung der Behandlungsdaten von rund 70 Millionen Bürger*innen in einer zentralen Forschungsdatenbank.[4]

Neu dabei ist, dass den Betroffenen kein Widerspruchsrecht eingeräumt wird, denn bislang galt für die Teilnahme an Studien immer die Einwilligung als oberstes Prinzip. Datenschutzbedenken begegnet Gesundheitsminister Jens Spahn mit dem Hinweis, dass die gespeicherten Informationen pseudonymisiert an die Sammelstelle – den Spitzenverband Bund der Krankenkassen – weitergeleitet würden. Tatsächlich werden zwar Name und Versichertennummer aus den Daten entfernt, sie enthalten jedoch weiterhin Alter, Geschlecht, Wohnort der Patient*innen sowie Daten zu den behandelnden Ärzt*innen.

Im Zeitalter von Big Data und Künstlicher Intelligenz zeigen sich immer mehr die Grenzen der Pseudonymisierung zum Schutz unserer sensiblen Informationen. Solche Datensätze erlauben in vielen Fällen noch Rückschlüsse auf Einzelpersonen – und zwar nicht nur auf die in dem Datensatz enthalte-

4 Bundesgesetzblatt Teil I, Nr. 49 vom 18.12.2019, S. 2562.

nen Personen selbst, sondern auch auf Unbeteiligte. Dabei lassen sich zwei Verfahrensweisen unterscheiden: zum einen die „Re-Identifikation" von Einzelpersonen in anonymisierten Datensätzen, zum anderen statistische Abschätzungen eigentlich schützenswerter Informationen anhand von „prädiktiven Analysen".

Der blinde Fleck unserer »Privatheit«

Das Risiko, in einem anonymisierten Datensatz vieler Patient*innendaten re-identifiziert zu werden, ist in der Informatik bereits seit Langem bekannt. Den Anstoß hierfür gab in den 1990er Jahren ein Fall aus den USA. Der US-Bundesstaat Massachusetts hatte damals die medizinischen Behandlungsdaten von rund 135 000 staatlichen Bediensteten und ihren Familienmitgliedern pseudonymisiert in einer Datenbank zu Forschungszwecken zusammengetragen. Dennoch gelang es der damaligen Informatikstudentin Latanya Sweeney durch das Kombinieren dieser Daten mit öffentlich zugänglichen Informationen aus dem Wähler*innen-Register von Massachusetts, die Krankenakte des damaligen Gouverneurs von Massachusetts, William Weld, zu rekonstruieren.[5]

Die spektakuläre Aktion sorgte für erhebliches Aufsehen und hat die Debatte um den Datenschutz in den Vereinigten Staaten stark geprägt. In der Informatik gilt Sweeneys Vorgehen als Musterbeispiel für einen Angriffstyp, der Informationen aus anderen zugänglichen Quellen heranzieht, um Daten auf diese Weise zu re-identifizieren. In der mathematischen Theorie der Datenbanksicherheit ist das Prädikat „anonym" heute daher nicht mehr gleichbedeutend mit sicher, sondern stellt vielmehr ein stark vom Kontext abhängiges Kriterium dar.

Eine zweite Form des Datenmissbrauchs ist derzeit aber noch virulenter, obschon sie in der öffentlichen Debatte weniger prominent ist: Mit Hilfe großer – potentiell anonymisierter – Datensätze lassen sich sogenannte prädiktive Analysen und Risiko-Scores erstellen, die ebenfalls sensible Informationen über Einzelpersonen offenlegen können. Sogenannte Korrelationsanalysen stellen dabei durch maschinelle Lernverfahren statistische Zusammenhänge zwischen privaten Informationen – etwa Krankheitsbefunden, psychologischen Behandlungen oder erblichen Vorbelastungen – und Verhaltensdaten her. Letztere fallen zum Beispiel bei der Nutzung von Fitness-Trackern, Smart Watches oder von sozialen Netzwerken an.

Auf diese Weise haben Mediziner*innen der University of Pennsylvania beispielsweise Postings auf Facebook daraufhin auswerten können, ob Nutzer*innen an Depressionen, Psychosen, Diabetes oder Bluthochdruck leiden. Und auch Facebook selbst setzt nach eigenen Angaben bereits seit Längerem KI-Systeme ein, um selbstmordgefährdete Nutzer*innen anhand ihrer Postings zu erkennen. Andere Unternehmen nutzen Kreditkartentransaktionen dazu, um Schwangerschaften bei ihren Kundinnen zu erkennen, um ihnen spezielle Werbung zuzustellen.[6]

Das Beispiel prädiktiver Analytik zeigt, dass Anonymisierung keine Gewähr für einen ausreichenden Datenschutz darstellt, sobald nicht nur die Daten einzelner Personen, sondern aggregierte Datensätze über Millionen von Patient*innen (Big Data) im Spiel sind. Denn prädiktive Analysen zielen nicht darauf ab, die Identität einer Person in den anonymisierten Daten aufzudecken. Vielmehr nutzen sie die gro-

5 Vgl. Latanya Sweeney, Weaving Technology and Policy Together to Maintain Confidentiality, in: „The Journal of Law, Medicine & Ethics", 25/1997, S. 98-110.

6 Vgl. Carles Duhigg, How Companies Learn Your Secrets, www.nytimes.com, 16.2.2012.

ße Menge verfügbarer Datensätze, um darin Muster zu erkennen und die Individuen in Risikogruppen einzuteilen. Diese Vorgehensweise ermöglicht es, sensible Informationen zum Beispiel über Krankheiten vorherzusagen – und zwar auch dann, wenn jemand selbst gar nicht in die Verarbeitung seiner Gesundheitsdaten eingewilligt hat.[7] Bewirbt man sich etwa auf einen Kredit oder eine Versicherung, können solche Analysen im Hintergrund ablaufen. Bei einem höheren geschätzten Risiko werden den Bewerber*innen dann schlechtere Konditionen angeboten.

Datenschutz in Zeiten von Big Data

Im Zeitalter von Big Data kann die eigene Privatsphäre somit durch Daten verletzt werden, die Millionen *anderer* Menschen über sich preisgeben – denn erst der millionenfache Vergleich mit den anonymen Daten anderer ermöglicht eine prädiktive Analyse. Diese Tatsache bildet den blinden Fleck unseres individualistischen Denkens über Datenschutz: Datenschutz wird im westlichen Diskurs in der Regel mit dem Recht jedes Einzelnen verbunden, die Speicherung und Verwendung seiner personenbezogenen Daten zu kontrollieren (informationelle Selbstbestimmung). Sobald ein Datensatz keine identifizierenden Informationen mehr enthält, weil er beispielsweise anonymisiert wurde, sehen die meisten Menschen in seiner Verwendung keine Gefahr mehr für sich selbst und willigen oft sogar in die Verwendung dieser Daten ein.

Gerade die Nutzung von Gesundheitsdaten zeigt, dass sich der Datenschutz im Zeitalter von Big Data und Künstlicher Intelligenz von der liberalistischen Konzeption der eigenen Privatsphäre lösen muss. Wir benötigen ein kollektivistisches Verständnis von Privatsphäre, welches berücksichtigt, dass auch anonymisierte Daten durch Mustererkennung in vielen Fällen dazu verwendet werden können, Einzelpersonen zu benachteiligen – und zwar nicht nur die in dem Datensatz enthaltenen Personen selbst, sondern auch Unbeteiligte. Die Folgen können mitunter dramatisch sein, etwa bei der Frage, ob man als Sicherheitsrisiko eingestuft und präventiv von der Polizei beobachtet oder bei einem Bewerbungsverfahren berücksichtigt wird.

Dagegen ist selbst die vergleichsweise fortschrittliche europäische Datenschutzgrundverordnung weitestgehend wirkungslos, da ihre Mechanismen gegen die Nutzung anonymisierter Daten in automatisierten Entscheidungen in der Praxis leicht auszuhebeln sind.[8]

Um den Missbrauch prädiktiver Analytik – nicht zuletzt im Gesundheitssektor – zu verhindern, benötigen wir daher ein erweitertes Verständnis von Privatsphäre, das sich nicht nur auf explizit *erhobene* Informationen, sondern auch auf *abgeschätzte* Informationen erstreckt. Zugleich bedarf es eines breiteren Bewusstseins darüber, dass prädiktive Analysen nur deshalb möglich sind, weil viele Bürger*innen sowie politische Entscheidungsträger*innen keine Einwände dagegen erheben, ihre Daten – auch anonymisiert – zur Verfügung zu stellen. Denn prädiktive Algorithmen, die *andere* Individuen als Abweichler*innen erkennen, lassen sich nur dann trainieren, wenn ausreichend viele Daten „normaler" Nutzer*innen vorliegen, die zumeist der Überzeugung sind, nichts zu verbergen zu haben. Das Gegenteil aber ist der Fall.

7 Zu den Risiken vgl. Virgina Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, New York 2017 sowie: Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York 2016.

8 Vgl. Sandra Wachter, Data Protection in the Age of Big Data, in: „Nature Electronics", 2/2019, S. 6-7.