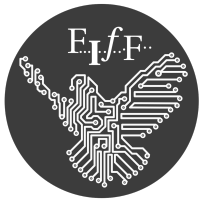

Stellungnahme zum

Datentransparenzverfahren

27. August 2022

Dirk Engling, Jens Kubieziel, Rainer Rehak

Diese Stellungnahme des Chaos Computer Clubs (CCC) und des Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FIfF) beschäftigt sich mit dem Konzept des Datentransparenzverfahrens aus technischer Sicht. Bestehende Implementationen standen zur Prüfung nicht zur Verfügung und sind nicht Teil dieser Stellungnahme.

1. Kurzbeschreibung des Verfahrens

Das Datentransparenzverfahren bildet die Grundlage für die Zusammenführung und Übermittlung von Gesundheitsdaten aller gesetzlich Versicherten. Im Rahmen des Verfahrens und des zugehörigen Digitale-Versorgungs-Gesetzes sollen diese Gesundheitsdaten zwangsweise in einer zentralen Datenbank gespeichert werden.

Die Daten sollen für Forschung und Gesundheitsberichterstattung zur Verfügung gestellt werden. Weiterhin sollen gesetzliche Kassen mit den Daten Planung, Analysen und Evaluationen durchführen können.

Folgende Angaben werden als Gesundheitsdaten verarbeitet:

- Geburtsjahr,
- Geschlecht,
- Postleitzahl (falls vorhanden),
- Sterbedatum, eventueller Vitalstatus[1],
- Daten zum Versicherungsverhältnis,
- Kosten- und Leistungsdaten (typischerweise Diagnosen, Behandlungen, Arzneimittel oder Krankengeld-Informationen).

[1] die Überlebenszeit, berechnet ab der gesicherten Diagnosestellung einer schweren Erkrankung.

Je Datensatz wird ein jährlich wechselndes Lieferpseudonym vergeben. Die Zusammenführung erfolgt zentral beim Spitzenverband Bund der Krankenkassen. Er übermittelt die gesamten Daten an das Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Die Datenaufbereitungsstelle des BfArM übernimmt hier die Rolle des Forschungsdatenzentrums: Es soll die Erschließung der Abrechnungsdaten aller gesetzlich Versicherten ermöglichen.

Bei der Zusammenführung wird lediglich das Lieferpseudonym aus jedem Datensatz entfernt und dadurch nicht übermittelt. An seine Stelle tritt eine Arbeitsnummer pro Datensatz, aus der das Lieferpseudonym nicht rückschließbar sein soll. Der Spitzenverband Bund der Krankenkassen übermittelt dann eine Lister aller Lieferpseudonyme mitsamt den neu dazu vergebenen Arbeitsnummern an das Robert-Koch-Institut (RKI), welches als Vertrauensstelle agiert.

Das RKI berechnet aus den Lieferpseudonymen nun dauerhafte Pseudonyme, die an die jeweiligen Patienten und deren Datensätze gebunden sind. Diese bleiben auch bei einem Wechsel der Krankenkasse gleich und sollen Langzeitforschung ermöglichen. Die dauerhaften Pseudonyme werden dann zusammen mit den Arbeitsnummern vom RKI an das BfArM übermittelt. Die Vertrauensstelle RKI soll danach die alle drei Arten von Pseudonymen löschen (Lieferpseudonyme, Arbeitsnummern und dauerhafte Pseudonyme).

2. Zentrale Speicherung

Ziel des Datentransparenzverfahrens ist die Zugänglichmachung der Gesundheitsdaten und der Krankengeschichte aller Patienten in Deutschland. Dies verlangt den bestmöglichen Schutz der feingranularen Daten der Krankenversicherten. Dass dabei

konzeptuell auf das Prinzip der Zentralisierung der Daten gesetzt wurde, steht diesem Ziel aus Sicht der Datensicherheit entgegen.

Gesundheitsdaten sind personenbezogene Daten mit einem sehr hohen Schutzbedarf. Der potentielle Eingriff in Rechte und Freiheiten der Versicherten ist also im Falle einer Verletzung der IT-Sicherheit besonders hoch. Die DSGVO sieht für Gesundheitsdaten ebenfalls einen hohen Schutzbedarf.[2] Im Vergleich zu anderen personenbezogenen Daten haben diese Daten auch auf dem Schwarzmarkt einen besonders hohen Wert. Diese Betrachtungen legen nahe, dass die Sicherheit der Daten einen hohen Stellenwert besitzen muss.

Eine zentrale Speicherung bietet für Angreifer den Vorteil, dass diese bei einem erfolgreichen Angriff den Zugriff auf alle Daten gelangen können. Somit gleicht aus Sicht der Betreiber und betroffenen Personen ein erfolgreicher Angriff einem Totalschaden. Eine dezentrale Speicherung vermindert solche Auswirkungen deutlich, da hier nur Teile des gesamten Datenbestandes in die Hände von Angreifern gelangen können. Zugleich erhält der Betreiber bei der Erkennung Wissen um den Angriff und kann diesen bei den anderen Teilen abwehren.

Auch hinsichtlich Fehlbedienungen ist ein zentrales System von Nachteil. Fehlbedienungen können sich beispielsweise auf den kompletten Datenbestand auswirken. Notwendige Korrekturen erfordern hohen Aufwand und hohe Kosten. Ein dezentrales System reduziert prinzipbedingt die Auswirkungen fehlerhafter Bedienung deutlich.

Vorgesehen ist die zentrale Datensammelstelle beim Spitzenverband Bund der Krankenkassen. Doch jede zentralisierte Lösung lädt nicht nur Angreifer ein, sondern bedeutet eine hier unnötige Machtkonzentration beim Betreiber selbst, etwa für den

---

[2] EG 54 DSGVO.

erweiterten Gebrauch, also Missbrauch, der Daten. Der Verlust von zentralisierten großen Datenmengen ist zudem besonders kostspielig. Wenn die Gesundheitsdaten an einen Angreifer verlorengehen, entstehen massive Probleme auf Seiten der gesetzlich Versicherten.

Aber auch auf Seiten der Datensammelstelle sind hohe Kosten durch juristische Verfahren, Schadensersatzforderungen und andere Folgekosten zu erwarten. Dezentral gespeicherte Daten haben den Vorteil, dass die Auswirkungen eines erfolgreichen Angriffs und somit die Anzahl der betroffenen Personen und die Kosten eines Datenverlusts schon ansatzbedingt viel geringer sind. Sofern also keine Notwendigkeit besteht, muss auf eine zentrale Speicherung verzichtet werden.

Weniger attraktiv für Angreifer sind Systeme ohne zentrale Instanz, bei der die Daten nicht an einem Punkt zusammenlaufen. Das hat den Vorteil, dass nicht allein auf die Kompetenz, das Wohlmeinen und die Achtsamkeit des Betreibers vertraut werden muss. Die Komplexität eines zentralen Systems ist größer und es ist viel attraktiver.

Die Argumente, die regelmäßig für eine zentrale Speicherung vorgebracht werden, basieren meist auf ökonomischen Betrachtungen bei Betrieb und auch Sicherung großer Datenmengen: Erhofft werden Skaleneffekte, vor allem sinkende Kosten. Allerdings können die Daten in einem dezentralen System so gut gesichert sein, dass es sich für einen Angreifer nicht lohnt, die hohen Hürden zu überwinden, weil die potentielle Beute so klein ist. Die dezentralen Datentöpfe im Verhältnis zu den Schutzmaßnahmen so klein zu machen, dass sich ein Angriff nicht lohnt, wäre daher ein erfolgversprechender Ansatz.

3. Pseudonymisierung und zu lange Speicherdauer

Oberstes Ziel der Verschleierung der zentralisierten Datensätze ist der Schutz der betroffenen Personen, insbesondere vor Re-Identifizierung und Datenmissbrauch. Als Hauptschutzmechanismus ist beim Datentransparenzverfahren konzeptuell die Pseudonymisierung vorgesehen.

Es sollen dabei verschiedene Schritte der Pseudonymisierung stattfinden. So wird einerseits ein Lieferpseudonym sowie eine Arbeitsnummer erzeugt, andererseits errechnet das RKI ein periodenübergreifendes, also dauerhaftes Pseudonym. Das Lieferpseudonym wird jährlich gewechselt.

Von einer wirksamen Pseudonymisierung wird aus technischer Sicht dann ausgegangen, wenn der Datensatz auch unter Hinzunahme zusätzlicher Daten nicht ohne weiteres einer konkreten Person zuzuordnen ist. Je mehr Daten für eine erfolgreiche Re-Identifizierung nötig sind, umso besser ist die Pseudonymisierung. So wäre eine Liste von Personalnummern einer Firma für eine außenstehende Person weitgehend wertlos. Eine Liste von Personalnummern und zugehörigen Anschriften ließe durchaus Rückschluss auf einzelne Personen zu.

Allerdings kann die Bewertung einer Pseudonymisierung nicht abstrakt, sondern immer nur konkret für einen Sachverhalt durchgeführt werden. Deshalb muss die Frage gestellt werden, inwieweit eine Pseudonymisierung bei Gesundheitsdaten überhaupt ein gutes Mittel zur Verhinderung von Re-Identifizierung ist. Denn Gesundheitsdaten sind in der Regel nur mit Kontext sinnvoll, sowohl bei der Nutzung durch Forscher als auch bei der missbräuchlichen Verwendung.

Ein Datensatz bezüglich einer Krankheit ist in der Regel nur dann nützlich, wenn auch Geburtsjahr, Geschlecht, Krankengeschichte usw. des Falles zugeordnet sind. Medizinische Studien arbeiten daher in der Regel so, dass Kontextdaten zu allen Probandinnen erhoben werden.

Allerdings sind es genau diese medizinisch notwendigen Zusatzdaten, die eine Re-Identifizierung des Datensatzes verhältnismäßig einfach möglich machen. Es ist im Zweifel für die Re-Identifizierung eher unerheblich, ob am Datensatz noch ein echter Name hängt oder ein konstruiertes Pseudonym. Der Datensatz beschreibt eine Person also in der Regel hinreichend, weswegen er ja überhaupt erst für die medizinische Forschung sinnvoll ist. Damit ergeben sich grundsätzliche Zweifel, ob die Pseudonymisierung überhaupt ein sinnvoller Schutzmechanismus gegen Re-Identifizierung sein kann.

Welche Schutzfunktion welche verschiedenen Pseudonymisierungsarten entfalten könnten, übersteigt den Umfang dieser Stellungnahme.[3] Die im Datentransparenzverfahren geplante simple Pseudonymisierung mit Arbeitsnummern und Lieferpseudonymen bei weiterhin gänzlich intaktem Restdatensatz ist jedoch prinzipiell keinesfalls der Aufgabe gewachsen, die zentralisiert gespeicherten Gesundheitsdaten von Millionen von Menschen adäquat zu schützen.

Je mehr die Daten jedoch mit hochkomplexen Pseudonymisierungsverfahren verschleiert werden, um Personen vor Re-Identifizierung zu schützen, umso schwieriger wird es, die Wahrnehmung von Datenschutz-Betroffenenrechten (beispielsweise Löschung, Berichtigung, Widerspruch) zu ermöglichen. Denn auch wenn personenbezogene Daten pseudonymisiert werden, so bleiben es personenbezogene Daten und ihre Verarbeitung verstößt potentiell gegen das Recht auf Datenschutz der Betroffenen.

---

[3] Vgl. Gutachten von Prof. Dr. Dominique Schröder.

Beispielsweise müsste die Person für ein Löschgesuch zunächst korrekt identifiziert und dann der zugehörige Datensatz gefunden werden – genau die individuelle Verknüpfung, die eigentlich mit allen Mitteln verhindert werden sollte. Dieser Widerspruch ist unauflösbar und zeigt, dass ein wirksamer IT-Schutzmechanismus eigentlich noch vor der Datenzentralisierung ansetzen müsste.

Bei dieser jetzt schon komplexen und auch perspektivisch unabsehbaren Gemengelage technischer, medizinischer und rechtlicher Schwierigkeiten ist eine Risikoabschätzung für eine Speicherdauer von dreißig Jahren gar nicht sinnvoll möglich, weder für die Individuen (Einwilligung/Widerspruch) noch für den Gesetzgeber selbst. Schon deswegen muss die Speicherdauer deutlich reduziert werden.

3. Weitergabe der Daten

Das Hauptziel des Datentransparenzverfahrens ist die Nutzung der Gesundheitsdaten, einerseits innerhalb der gesetzlichen Krankenkassen zur Planung, Analyse und Evaluation der Patientinnenversorgung und andererseits durch externe Nutzungsberechtigte. Dabei ist das BfArM als Forschungsdatenzentrum Drehscheibe und Verteilungszentrum. Um also die Zwecke der medizinischen Forschung, Versorgungsforschung und auch Gesundheitsberichterstattung zu verfolgen, sollen die zentral gespeicherten pseudonymisierten Gesundheitsdaten nach positiv beschiedenem Antrag weitergegeben werden dürfen.

Das Forschungsdatenzentrum hat nach der Übermittlung der Datensätze selbst keine Einflussmöglichkeit mehr auf die übermittelten Kopien. Auch Behördenhandlungen und rechtliche Schritte können die Datenverarbeitung, etwa bei Falschübermittlung, nur innerhalb der deutschen Jurisdiktion gerade noch sinnvoll kontrol-

lieren. Darum sind besonders strenge Anforderungen hinsichtlich des Datenschutzes und der IT-Sicherheit absolut nötig, sowohl im Allgemeinen als auch für konkrete Auswertungsprojekte. Mindestens müssen zwei Fragen klar beantwortbar sein: Welche Daten können zusammengeführt, übermittelt und anderweitig verarbeitet werden? Wie werden unautorisierte Zugriffe verhindert? Beide Fragen werden im geplanten Datentransparenzverfahren bislang nicht hinreichend geregelt.

Hinzu kommt ein wesentlicher Aspekt des realen Forschungsalltags. Wissenschaftliche Institute in Deutschland werten Daten nicht immer selbst aus, sondern nutzen dafür Dienste zur Datenauswertung. Diese bieten statistische Analysen sowie „Big-Data“-Auswertung oder Künstliche-Intelligenz-Systeme auf Basis von elaborierten Machine-Learning-Algorithmen an. Deren Hauptgeschäftsfeld anwendungsseitig ist momentan das Wiedererkennen von Besuchern unterschiedlicher Webseiten – mit anderen Worten das De-Pseudonymisieren von Personen als potentiellen künftigen Werbeempfängern. Mustererkennung auf vergleichsweise akribisch gepflegten Gesundheitsdaten stellt für diese Anbieter faktisch keine Hürde dar. Datenverarbeiter haben bestimmte Stammdaten vieler Menschen zudem schon vorrätig.

Auch wenn die Daten in der Regel vor Weitergabe anonymisiert und aggregiert werden, zeigen die obigen Überlegungen, dass eine vollständige unauflösliche Anonymisierung technisch schwer möglich und sachlich oft unbrauchbar ist. Für einige bestimmte Forschungszwecke ist zudem die Übermittlung der pseudonymisierten Einzeldatensätze möglich.

Diese Art von aussagekräftigen Gesundheitsdaten könnten folglich an Auswertungsdienste außerhalb deutscher oder europäischer Jurisdiktion übergeben werden, also eine Übermittlung an Drittstellen in Drittstaaten. Theoretisch und praktisch könnten diese Gesundheitsdatensätze dann aus verschiedenen Anfragen an verschiedene

Forschungsdatenauftragsverarbeiter im Ausland zusammengeführt und wieder zu einem größeren Datensatz zusammengefasst werden. Auch aus anonymisierten und aggregierten Daten lassen sich viele Rückschlüsse ziehen – besonders über eine lange Zeit. Da die Daten nach bisheriger Planung bis zu dreißig Jahre lang vorgehalten werden sollen, werden bestimmte Datensätze vermutlich Gegenstand diverser Übermittlungen sein und somit auch wiedererkennbar.

Am Beispiel von Auswertungsdiensten mit Bezug zur USA (also Firmen in den USA oder US-Firmen in Deutschland) zeigt sich, wie real die Gefahr der Übermittlung an Drittstellen sein kann. Nicht erst seit den Enthüllungen von Edward Snowden[4] wurde bekannt, dass die in der EU datenschutzrechtlich gebotenen und offiziell gemachten Zusagen mit herrschendem US-amerikanischen Recht nicht praktisch einhaltbar sind. Staatliche Stellen – etwa Geheimdienste – können unter bestimmten Umständen schrankenlos auf beliebige Daten der Nicht-US-Bevölkerung zugreifen, etwa auf Gesundheitsdaten der deutschen Bevölkerung. Mit dieser Begründung wurde mit dem Schrems-I-Urteil des Europäischen Gerichtshofs bereits im Jahre 2015 das Safe-Harbor-Abkommen zwischen der EU und den USA gekippt. Auch der Nachfolger „EU-US-Privacy-Shield“ hatte aus den gleichen Gründen wenig Bestand und wurde mit dem Schrems-II-Urteil des EuGH im Jahre 2020 für ungültig erklärt.

Eine Nachfolgeregelung besteht bisher nicht. Ihr kann jedoch ein ähnlich kurzes Leben vorhergesagt werden, solange sich die US-amerikanische Gesetzeslage nicht ändert. Am CLOUD-Act von 2018, wonach US-Firmen bestimmten US-Behörden den Zugriff auf Daten gewähren müssen, selbst wenn die Daten gar nicht in den USA verarbeitet werden, lässt sich jedoch ablesen, dass eine Änderung bislang nicht in Sicht ist.

---

[4] Vgl. https://www.cjfe.org/snowden

3. Schlussfolgerung

Aufgrund der Sicherheitsrisiken muss allen von der Datensammlung betroffenen Menschen ein Wahlrecht gelassen werden, ob sie sich daran beteiligen wollen oder nicht. Unabhängig davon ist das Sicherheitsniveau der gespeicherten Daten deutlich zu erhöhen. Damit kann Forschung ermöglicht, aber die Daten der versicherten Personen zugleich bestmöglich geschützt werden.

Die bisher vorgesehene sehr lange Speicherdauer von bis zu dreißig Jahren birgt erhebliche Risiken für die höchst sensiblen personenbezogenen Daten, denen das Sicherheitskonzept nicht ausreichend gerecht wird.

Bei Betrachtung der beabsichtigten Ziele des Datentransparenzverfahrens stellt sich die Frage, ob es nicht einfachere und weniger riskante Wege gäbe, um wissenschaftliche Forschung mit Gesundheitsdaten zu ermöglichen. Hier drängt sich sofort der Gedanke an eine Verpflichtung aller Krankenkassen auf, Wissenschaftlern über eine genormte Datenverarbeitungsschnittstelle Zugriff auf die in ihren Datenzentren sowieso schon vorgehaltenen und verarbeiteten Patientendaten zu gewähren, um beispielsweise statistische Auswertungen anzustoßen. Ein zentrales Gremium kann dann über die Vergabe von Zugangserlaubnissen für bestimmte Forschungen entscheiden. Die meisten der in dieser Stellungnahme aufgeworfenen Probleme würden sich bei einem solchen Ansatz gar nicht erst stellen.