Kay Hamacher, Tobias Kussel, Tatiana von Landesberger, Tom Baumgartl, Markus Höhn, Simone Scheithauer, Michael Marschollek, Antje Wulff

Fallzahlen, Re-Identifikation und der technische Datenschutz

Ein Aspekt der SARS-CoV-2-Pandemie

Die SARS-CoV-2-Pandemie hat viele sehr spezielle Fragen des Datenschutzes und der Datensicherheit aufgeworfen. Der vorliegende Beitrag widmet sich den mit der Veröffentlichung von Infiziertenzahlen verbundenen Re-Identifikationsrisiken. Er zeigt einen Weg auf, diese Risiken mit Mitteln des technischen Datenschutzes zu reduzieren, um sowohl das öffentliche Informationsbedürfnis zu befriedigen als auch das informelle Selbstbestimmungsrecht der Betroffenen zu wahren.

Prof. Dr. Kay Hamacher

ist Professor im Bereich der Bioinformatik und Simulation an der TU Darmstadt. Er arbeitet interdisziplinär in den Fachbereichen Biologie, Informatik und Physik an computer-gestützter Forschung, u.a. zum technischen Datenschutz in der Medizininformatik und personalisierten Medizin.
E-Mail: kay.hamacher@tu-darmstadt.de

Tobias Kussel, M.Sc.

ist Physiker an der TU Darmstadt. Seine Forschungsschwerpunkte sind Privatsphäreschutz durch kryptographische Protokolle, Genomic Privacy und Dynamik komplexer Graphensysteme.
E-Mail: tobias.kussel@tu-darmstadt.de

Prof. Dr. Tatiana von Landesberger

Lehrstuhl für Visualisierung und Visual Analytics, Mathematisch-Naturwissenschaftliche Fakultät, Universität zu Köln

E-Mail: visva.kontakt@cs.uni-koeln.de

Tom Baumgartl

Fraunhoferinstitut für Graphische Datenverarbeitung Darmstadt, Graphische Interaktive Systeme

E-Mail:
tom.baumgartl@gris.tu-darmstadt.de

Einleitung

Seit März 2020 befinden wir uns laut der Weltgesundheitsorganisation (WHO) in einer globalen Pandemie des Virus SARS-CoV-2. Diese Lage hat zu vielen neuen Fragen des Datenschutzes und der Datensicherheit geführt, von denen einige bereits in der DuD diskutiert wurden, beispielsweise die Problemstellungen der Impf- und Testnachweise (Schrahe und Städter 2021), des digitalen Schulunterrichts (Smoltczyk 2021) oder der Kontrolle und Überwachung von Beschäftigten (Dietrich, Bosse und Schmitt 2021) – auch und besonders im Home-Office.

In einem früheren Beitrag (Hamacher u. a. 2020) stellten wir einige Aspekte des technischen Datenschutzes genomischer Daten vor. In dem hier vorliegenden Beitrag geht es um die Veröffentlichung der Infiziertenzahlen, Re-Identifikationsrisiken und einen beispielhaften Weg, die so entstehenden Probleme mit Mitteln des technischen Datenschutzes zu reduzieren oder gar zu lösen. Dadurch werden sowohl das öffentliche Informationsbedürfnis befriedigt als auch das informelle Selbstbestimmungsrecht der einzelnen Betroffenen gewahrt.

1 Spannungsfeld Datenschutz und Informationsbedürfnis

Gerade zu Beginn der SARS-CoV-2 Pandemie und vor den traditionellen Familienfeiertagen Weihnachten und Ostern wurde der Konflikt zwischen zwei grundlegenden Interessen beobachtbar: Datenschutzinteressen der einzelnen Bürger auf der einen Seite mussten mit dem Informationsbedürfnis der Öffentlichkeit sowie der Politik abgewogen werden. Leider zeigten sich in dieser außergewöhnlichen Situation regulatorische Datenschutzmecha-

nismen als problematisch; die Veröffentlichung von spezifisch regionalen Fallzahlen war allgegenwärtig. Auch wenn dieses Vorgehen nach Art. 85 Abs. 2 Datenschutzgrundverordnung (DSGVO) prinzipiell möglich ist, wurden damit die Betroffenen sorglos dem Risiko der Re-Identifikation ausgesetzt, wenn nicht sogar die genauen Personaldaten z. B. an Polizeibehörden weitergegeben wurden (Laufer 2020).[1]

1.1 „Anonymisierte" Veröffentlichung der Infiziertenzahlen

Es ist offensichtlich, dass die ungeschützte, kleinteilige Veröffentlichung der Anzahl der mit SARS-CoV-2 infizierten Personen nicht mit den informellen Schutzrechten der Betroffenen vereinbar ist. Die Einführung und Bewertung geeigneter Schutzmaßnahmen (Art. 32 Abs. 2 DSGVO) erweist sich jedoch als schwierig: So muss nicht nur bei der Abwägung berücksichtigt werden, dass die Positiv/Negativ-Zählung der SARS-CoV-2 Tests als „Gesundheitsdaten" im Sinne des Erwägungsgrundes 35 der DSGVO und § 46 Pkt. 13 des Bundesdatenschutzgesetzes zählen und damit besonders schutzwürdig sind, auch die Wirksamkeit der ergriffenen Maßnahmen an sich lässt sich selten messen. Dies zeigt sich bereits in der uneinheitlichen Interpretation der Rahmenbedingungen, wie zum Beispiel der Diskrepanz zwischen der relativen und absoluten Bestimmbarkeit des Personenbezugs (Klar und Kühling 2018).

Die gängige Praxis die Fallzahlen als „anonymisierte" Daten zu veröffentlichen beantwortet die Datenschutzbedenken, da für anonymisierte Daten keine Verarbeitungserlaubnis nach Art. 6 Abs. 1 DSGVO notwendig ist und diese nicht dem Datenschutzrecht unterliegen. Im Folgenden soll es um die Frage gehen, ob die gängigen Anonymisierungsmethoden dem Schutz dieser Gesundheitsdaten angemessen sind.

k-Anonymität und Re-Identifikations-Angriffe

Datensätze gelten als anonymisiert, wenn sie derart verändert sind, dass die „Einzelangaben über persönliche oder sachliche Verhältnisse" (§ 3a BDSG a.F. 2009) nicht oder nur mit unverhältnismäßigem Aufwand einer Person zugeordnet werden können. Damit ist nun auch die Identifikation über sogenannte „Quasi-Identifier" eingeschlossen, also Daten, die nicht *per se* als identifizierend gelten, aber einzigartig genug sind, um in Verbindung mit anderen Quasi-Identifiern einer Person zugeordnet werden können.[2]

Die Anonymisierung stellt stärkere Anforderungen an die Daten als eine *Pseudonymisierung*. Bei dieser kann über die Hinzuziehung zusätzlicher Informationen der Personenbezug hergestellt werden (Art. 4 Abs. 5 DSGVO, § 46 Pkt. 5 BDSG). Ein Beispiel wäre eine Datenbank mit medizinischen Daten, bei denen die identifizierenden Daten durch eine eindeutige Patientenkennnummer ersetzt sind, vorausgesetzt, dass diese Datenbank und die Zuordnungstabelle zwischen Patientenkennnummer und identifizierenden Daten separat gespeichert werden.

Stellvertretend für die jeweiligen Landesregelungen werden hier die Anonymisierungsregeln des Landesbeauftragten für Datenschutz und die Informationsfreiheit Rheinland-Pfalz betrachtet.[3] Diese sind in den Corona-Datenschutz-FAQs der Behörde zu finden (Landesbeauftragter für den Datenschutz und die Informationsfreiheit Rheinland-Pfalz o. J.) und sind (nahe-

1 Die aktuellen Ereignisse und Ermittlungen rund um den unbefugten Zugriff auf persönliche Daten durch Polizeibeamte lässt diese Praxis in einem noch schlechteren Licht stehen. Im April 2020, dem Zeitpunkt der Datenübergabe, waren bereits Schutzmaßnahmen nach dem Infektionsschutzgesetz (IfSG) angeordnet, sodass eine Rechtmäßigkeit der Datenübermittlung gemäß § 19 Abs. 1 Satz 3 Landesgesetz über den öffentlichen Gesundheitsdienst Baden-Württemberg (ÖGdG-BW) angezweifelt werden kann.

2 In einer amerikanischen Studie (Sweeney 2000) wurde gezeigt, dass fast 90% der Datensätze des U.S.-Zensus von 1990 über die Merkmale „Postleitzahl", „Geschlecht" und „Geburtsdatum" eindeutig zu identifizieren sind. Mit den Merkmalen „Regierungsbezirk" (engl. County), „Geschlecht" und „Geburtsdatum" sind immerhin noch fast ein Fünftel der U.S.-Bürger eindeutig zu identifizieren. Diese Studie alleine, nun schon zwei Dekaden alt und in Datenschutzkreisen sehr bekannt, ist ein Grund, der Veröffentlichung von SARS-CoV-2-Fällen, tagesaktuell und nach Landkreisen aufgeschlüsselt, skeptisch gegenüberzustehen. Zusammen mit dem gerade zu Beginn der Pandemie großen Interesse an dem Altersprofil der Betroffenen lässt sich ein reelles Re-Identifikationsrisiko attestieren – ohne Korrelation mit weiteren Datenquellen.

3 Lesenswert ist auch der vom Landesbeauftragten für Datenschutz und die Informationsfreiheit Rheinland-Pfalz in der DuD veröffentlichte Artikel zu den durch die beschleunigte Digitalisierung entstandenen Herausforderungen der Behörde (Kugelmann 2021).

Markus Höhn

Fraunhoferinstitut für Graphische Datenverarbeitung Darmstadt, Graphische Interaktive Systeme

E-Mail: markus.hoehn@gris.tu-darmstadt.de

Prof. Dr. Simone Scheithauer

Universitätsmedizin Göttingen, Krankenhaushygiene & Infektologie,

E-Mail: krankenhaushygiene.leitung@uni-goettingen.de

Prof. Dr. med. Dr.-Ing. Michael Marschollek

geschäftsführender Direktor, Peter L. Reichertz Institut, Medizinische Hochschule Hannover

E-Mail: michael.marschollek@plri.de

Dr. Antje Wulff

Wiss. Mitarbeiterin, Peter L. Reichertz Institut, Medizinische Hochschule Hannover

E-Mail: antje.wulff@plri.de

zu) identisch mit den Regelungen anderer Bundesländer, geben allerdings, da sie zur Information der breiten Öffentlichkeit gedacht sind, auch Einblicke in die getroffenen Erwägungen und betrachteten Risiken.

Dort wird dargestellt, dass zum Schutz vor direkter Identifikation und der indirekten Identifikation über Quasi-Identifier einige Merkmale, wie das Alter, nur in kategorisierter Form und regional aggregiert zu veröffentlichen ist. Dabei ist die Stadt-/Verbandsgemeindeebene die tiefste zulässige regionale Gliederungsstufe. Für das Merkmal „Alter" würden beispielsweise alle Infektionszahlen eines Landkreises veröffentlicht werden, bei denen die Patienten zwischen 70 und 80 Jahre alt sind, anstatt das genaue Alter oder den genauen Wohnort zu beschreiben. Dies stellt eine Implementierung der sogenannten *k*-Anonymität dar.

_k_-Anonymität

Ein Datensatz erfüllt formal *k*-Anonymität (Sweeney 2002), wenn die identifizierenden Merkmale jedes Datums von mindestens *k* Einträgen im Datensatz erfüllt werden. Dabei wird je nach Datenschutzanforderung eine bestimmte natürliche Zahl für *k* gewählt. In der medizinischen Praxis gelten Daten oft als hinreichend anonymisiert, wenn eine *k*-Anonymität für $k = 5$, seltener auch $k = 11$ oder $k = 3$, erfüllt ist (European Medicines Agency 2017; Oswald 2013). Um *k*-Anonymität zu erreichen, werden die Datensätze in Äquivalenzklassen gruppiert. Im obrigen Beispiel könnte eine Äquivalenzklasse „A" für männliche Patienten im Alter zwischen 70 und 80 Jahren im Postleitzahlenraum 64XXX stehen.

Re-Identifikations-Angriffe

Das vorgestellte Anonymitätsmaß und weitere Metriken wie *l*-Diversität oder *t*-Closeness (Machanavajjhala u. a. 2007; Li, Li, und Venkatasubramanian 2007) sind konzeptionell einfach verständlich und einfach einzusetzen. Sie weisen jedoch, wenn auch in unterschiedlichem Ausmaß, Schwachstellen auf, die eine Re-Identifikation ermöglichen. Zudem ist die Abschätzung von praktikablen Parametern schwierig. So verlieren die Daten zum Beispiel bei Erhöhung des *k*-Parameters der *k*-Anonymität deutlich an Nutzbarkeit, da in immer gröbere Klassen sortiert werden muss, um die höhere Anonymität zu gewährleisten. Schließlich sind alle genannten Maße schwierig einzusetzen, falls nicht nur ein sensibles Merkmal, z. B. eine Diagnose, geschützt werden muss (Liu, Jia, und Han 2012; Wu u. a. 2010).

Die so genannte Linkage-Attacke auf anonymisierte Datensätze ist nicht nur akademischer Natur, sondern praktikabel und wird gegen anonymisierte Datensätze echter Produktivsysteme eingesetzt. Als Linkage-Attacke wird ein Re-Identifikationsversuch bezeichnet, bei dem die pseudonymisierten oder anonymisierten Datensätze mit weiteren dem Angreifer bekannten Datensätzen korreliert werden.

Das kanonische Beispiel einer Linkage-Attacke ist die erfolgreiche Reidentifikation des U.S.-Gouverneurs William Weld im Jahr 1997. Der Erzählung nach wurden Welds medizinische Daten in einem pseudonymisierten Versicherungsdatensatz mit den öffentlichen Wählerregistern von Cambridge, Massachusetts korreliert und damit Weld re-identifiziert. Systematische Untersuchungen (Barth-Jones 2012) zeigen jedoch, dass dieser Hergang wohl eher ein Mythos ist. Viel wahrscheinlicher hat die Verbindung der Versicherungsdaten mit der öffentlichen Krankenhauseinlieferung des Gouverneurs zur Re-Identifikation geführt – dennoch eine erfolgreiche Linkage-Attacke.

Das Beispiel illustriert einen fundamentalen Aspekt, der in Linkage-Attacken ausgenutzt wird: Ein Datenhalter kann nur die Pseudonymisierung und Anonymisierung seiner eigenen Daten kontrollieren. Die vollständige Deanonymisierung eines Datensatzes durch die Hinzuziehung neuer, zusätzlicher Datenquellen ist nie auszuschließen. Selbst wenn ein Datenhalter die Anonymisierung unter Einbeziehung aller zum Zeitpunkt der Veröffentlichung bekannten externen Datensätze, in dem ein Teil der „eigenen" Nutzer enthalten sein könnte, durchführt, besteht dieses Risiko. Beispielhaft kann hier die Erörterung von Ohm (Ohm 2010) zu externem Wissen genannt werden.

In der gegenwärtigen Situation werden nicht nur SARS-CoV-2 Fallzahlen auf Stadtebene oder für Landkreise aufgeschlüsselt veröffentlicht, sondern auch die Versorgungskapazitäten vieler Krankenhäuser aufgeteilt in reguläre und intensivmedizinische Versorgungskapazität. Dadurch lassen sich in kleineren Landkreisen die einzelnen SARS-CoV-2-Fälle spezifischen Krankenhäusern zuordnen. Weiterhin stellt die tagesaktuelle Veränderung der Zahlen und Kapazitäten ein großes Linkage-Risiko dar.

2 Sichere Fallzahlauswertung mittels Secure Multi-Party Computation

In diesem Abschnitt möchten wir eine konkrete Anwendung zeigen, die auf der Grundlage allgemeiner Konzepte des technischen Datenschutzes viele Analysen bei zeitgleicher Garantie hoher Datensicherheit erlaubt. Konkret basieren die Sicherheitsgarantien auf der nachfolgend vorgestellten kryptografischen Technik der *Secure Multi-Party Computation* (SMPC). Damit soll exemplarisch gezeigt werden, dass sich viele der Anforderungen an Datenveröffentlichung und -übermittlung im Kontext der SARS-CoV-2-Pandemie auch ohne erhebliche Datenschutzrisiken hätten erfüllen lassen.

Bei der gezeigten Anwendung handelt es sich um ein akademisches Projekt, jedoch gibt es mittlerweile genügend industrielle Expertise und Implementierungen, um den produktiven Betrieb außerhalb der Universitäten und Universitätsklinika möglich zu machen.

Wie wir bereits dargestellt haben, stellt die geographisch feingranulare Veröffentlichung der Fallzahlen ein erhebliches Datenschutzrisiko dar. Jedoch wären viele – auch für die breite Öffentlichkeit interessante – Auswertungen ohne Publikation der Zahlen möglich.

2.1 Secure Multi-Party Computation

In einer idealen Welt wären verteilte Berechnungen trivial zu lösen, indem die geheim zu haltenden Eingangsdaten einer absolut vertrauenswürdigen Instanz übermittelt würden, die die Berechnung durchführt und nur das Ergebnis veröffentlicht. In der realen Welt nimmt das Robert Koch Institut (RKI) diese Rolle in Bezug auf die SARS-CoV-2-Infektionszahlen ein. Diese zentralisierte Speicherung ist nicht ohne Datenschutzrisiken.

Mit den Techniken der Secure Multi-Party Computation wird diese ideale vertrauenswürdige Instanz durch ein kryptografisches Protokoll *simuliert*. Die Eingangsdaten der an der Berechnung beteiligten Parteien werden dabei jedoch nicht an die anderen Parteien übertragen, sondern in einem vom konkreten Protokoll abhängigen Verfahren encodiert. Dabei wird rigoros be-

weisbar sichergestellt, dass ein extrem hohes Schutzniveau, bis hin zur *informationstheoretischen Sicherheit*[4] bei einigen Protokollen, eingehalten wird.

Das Feld wurde 1982 von Andrew Yao mit der Formulierung und Lösung (Yao 1982) des *Millionaires' Problem*[5] begründet. Wenige Jahre später verallgemeinerten Yao (Yao 1986), sowie Goldreich, Micali und Wigderson (Goldreich, Micali, und Wigderson 1987) die Techniken mit den Protokollen „Yao's Garbled Circuits" und dem nach den Begründern benannten „GMW"-Protokoll auf beliebige Fragestellungen. Zum Zeitpunkt der Veröffentlichung waren die Protokolle jedoch rein theoretische Konstrukte; Limitationen der Rechen- und Übertragungstechniken verhinderten einen praktischen Einsatz. Erst mit der Entwicklung des „Fairplay"-Compilers im Jahr 2004 (Malkhi u. a. 2004) und durch Verbesserungen der Computertechnik und Optimierung der theoretischen Protokolle begann die praktische Nutzbarkeit generischer SMPC-Protokolle für einige Anwendungen. In ihrer weiterentwickelten und optimierten Form sind sowohl Yao's Garbled Circuits als auch das GMW-Protokoll immer noch relevant und im Einsatz (siehe z. B. Ashur u. a. 2021; Rosulek und Roy 2021; Braun u. a. 2020; Patra u. a. 2020).

Die praktische Anwendbarkeit ist meist durch die im Vergleich zu Klartextanalysen schlechte Performance limitiert; oft ist die Laufzeit der Analysen drei Größenordnungen oder mehr langsamer. Die Optimierung von geeigneten Anwendungen sowie die Optimierung der Protokolle selbst sind jedoch aktiver Gegenstand der Forschung, sodass immer mehr Anwendungsfälle, nachgewiesen durch wissenschaftliche Publikationen, in der Praxis möglich sind (siehe z. B. Addanki u. a. 2021; Agrawal u. a. 2020; Carpov u. a. 2021; Keller u. a. 2021; Sabelfeld 2019; Vogelsang u. a. 2020; Stammler u. a. 2020). In der hier besprochenen Anwendung sind allerdings die Laufzeiten der Algorithmen nicht nennenswert.[6]

2.2 Secure-Histogramm-Anwendung des HiGHmed Konsortiums

Im Kontext des Schwerpunktes „Infektionskontrolle" innerhalb des Konsortiums „HiGHmed" der Medizininformatik-Initiative des Bundesministeriums für Bildung und Forschung (BMBF) (Haarbrandt u. a. 2018) bearbeiten die Autoren das Thema Datenschutz sowie die Berechnung und Visualisierung von Infektionswegen. Aus aktuellem Anlass wurde ein System für die institutsübergreifende Berechnung von Histogrammen, d. h. nach bestimmten Kriterien aggregierte Werte, entwickelt. Für den Prototyp wurde die Darstellung von SARS-CoV-2-Fällen, gruppiert nach Blutgruppe, gewählt.[7] Als Datengrundlage wurden dabei synthetische Daten verwendet, deren Merkmale, wie die Blutgruppe, der realen demographischen Verteilung entsprechen. Die Berechnung erfolgt über ein spezialisiertes Secure-Multi-Party-Computation-Protokoll, welches durch Arithmetisches Secret-Sharing[8] und dessen additive Homomorphieeigenschaften informationstheoretische Sicherheit der Rohdaten garantiert. Die additive Homomorphie ist dabei das grundlegende Charakteristikum, welches das arithmetische Secret-Sharing für verteilte Histogrammberechnung über Aufaddieren der lokalen Fallzahlen so effizient macht. In der von uns implementierten Protokollversion können beliebig viele Standorte angeschlossen werden, solange die Summe der Fallzahlen $2^{127} - 1 \approx 1.7 \times 10^{38}$ nicht überschreitet.

Der Ablauf der Berechnung ist in Abbildung 1 schematisch gezeigt. In zwei Kommunikationsrunden werden Secret Shares ausgetauscht. Die weiteren Berechnungen finden lokal und ohne Mitwirkung der anderen Parteien statt.

Abbildung 1 | Schema des SMPC-Protokolls zur Berechnung eines Histogramms über verteilte Daten in zwei Interaktionsrunden. Zuerst wird die Berechnung von einem Standort initialisiert. Es erfolgt im zweiten Schritt das Versenden und Empfangen der zu den Eingabedaten gehörenden Secret Shares. Diese werden in Schritt drei lokal zu einem Ergebnis-Share summiert. In Interaktionsrunde zwei werden nun diese Ergebnis-Shares zwischen allen Parteien ausgetauscht. Nun kann im letzten Schritt das Ergebnis rekonstruiert und im Klartext angezeigt werden.

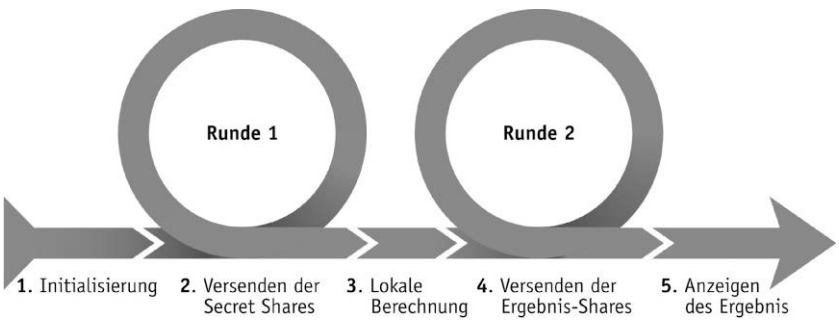


Abbildung 2 zeigt ein auf Basis von anonymisierten Fallzahlen erzeugtes Histogramm. Die Blutgruppeninformation ist dabei für Testzwecke synthetisch erzeugt, sodass das Ergebnis nicht der medizinischen Sachlage entspricht. Die Anwendung kann direkt in das SmICS-System[9] zur Infektionskontrolle eingebunden wer-

4 *Informationstheoretisch sicher* bedeutet, dass keinerlei Informationen über den Klartext übertragen werden, das Chiffrat ohne den dazu gehörigen Schlüssel also selbst von einem Angreifer mit unbeschränkter Rechenkapazität und -zeit nicht entschlüsselt werden kann. Es ist bei einem Entschlüsselungsversuch noch nicht einmal überprüfbar, ob eine korrekte Nachricht entschlüsselt wurde, da jede beliebige Nachricht mit der Länge des Chiffrats als Ergebnis der Entschlüsselung möglich ist.

5 Das *Millionaires' Problem* dreht sich um die Frage, wie der Vermögendste einer beliebigen Anzahl (fiktiver) Millionäre gefunden werden kann, ohne dass das Vermögen der Beteiligten preisgegeben werden muss – die erste veröffentlichte Multi-Party Computation.

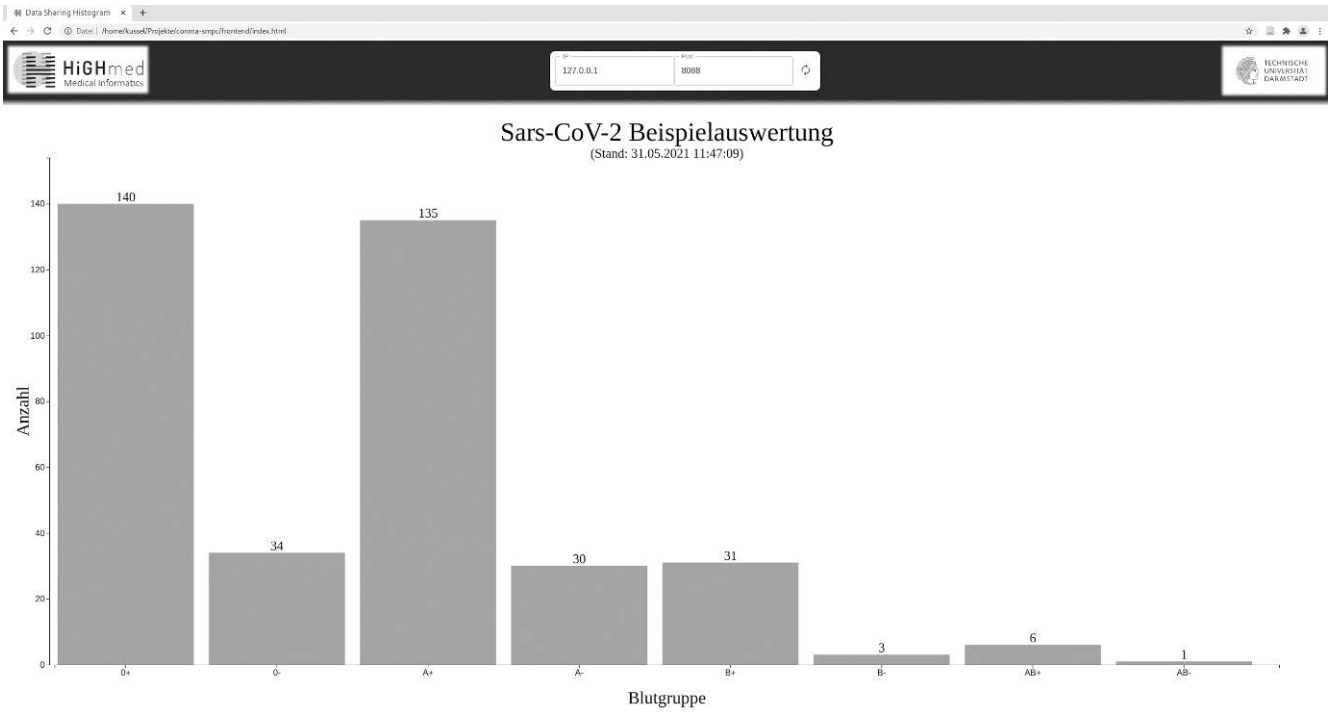
6 Die Laufzeit betrug in unseren Tests gemittelt über zehn Durchläufe bei drei beteiligten Parteien und a) 10.000 b) 100.000 c) 1.000.000 Patienten pro Partei je (4.17±0.12)s, (5.91±0.22)s, bzw. (16.07±1.05)s.

7 Diese Darstellung ist besonders interessant für die SARS-CoV-2-Forschung und -Versorgung, da neue Studien Abhängigkeiten der Symptomatik und der Infektionsrisiken vermuten (Ellinghaus, Degenhardt, Bujanda, Buti, Albillos, Invernizzi, Fernandez, u. a. 2020a; Ellinghaus, Degenhardt, Bujanda, Buti, Albillos, Invernizzi, Fernández, u. a. 2020b; Ellis 2020; Zietz und Tatonetti 2020).

8 Die Idee hinter arithmetischem Secret-Sharing ist, dass die geheimen Informationen durch das mehrmalige Mischen mit großen Zufallszahlen in mehrere Teile, sogenannte „Shares", aufgeteilt werden. Jeder Teil für sich enthält keine Information, insofern, dass er von einer reinen Zufallszahl nicht zu unterscheiden ist. Der geheime Wert kann über die Rekombination **aller** Shares zurückerhalten werden; fehlt nur ein einziges Share ist dies nicht möglich.

9 Das Smart Infection Control System (SmICS) wird zur Nachverfolgung und Visualisierung von Infektionsketten im Krankenhaus sowie zur frühzeitigen Erkennung von Infektionsclustern und -ausbrüchen durch das HiGHmed-Konsortium entwickelt. Seit Ende Mai 2020 wird das System auch für die SARS-CoV-2-Erforschung weiterentwickelt und bereits an mehreren Universitätskliniken

Abbildung 2 | Bildschirmfoto des durch die SMPC-Anwendung der TU-Darmstadt entstandenen SARS-CoV-2-Fall per Blutgruppen Histogramms. Für die Auswertung im Bild sind synthetische Daten verwendet worden, die nicht die tatsächliche medizinische Sachlage widerspiegeln. Während der Berechnung unterliegen die Rohdaten informationstheoretischer Sicherheit und genügen damit höchsten Datenschutzanforderungen.



den und erlaubt den angeschlossenen Universitätskliniken damit wichtige institutsübergreifende Auswertungen, ohne dabei die Privatsphäre der Patienten zu gefährden.

Wenn anstelle von „Blutgruppe" und „Anzahl" die Merkmale „Alle Tests" und „Positive Tests" gewählt werden, kann ohne weitere Umstellung des Systems auch ein Mittelwert wie zum Beispiel eine 7-Tages-Inzidenz berechnet werden.

Fazit

In diesem Beitrag haben wir dargestellt, dass die Anonymisierungsmittel, die zum Schutz der medizinischen Daten von mit SARS-CoV-2 infizierten Menschen verwendet werden, nicht ausreichend sind, um ein substanzielles Re-Identifikationsrisiko auszuschließen. Es existieren jedoch gut erforschte Mechanismen des technischen Datenschutzes, die diese Risiken erheblich verringern oder sogar komplett eliminieren. Dafür wurde beispielhaft eine Arbeit der TU Darmstadt zur Infektionskontrolle im Forschungsprojekt HiGHmed vorgestellt.

Die Autoren wünschen sich für die Zukunft einen verstärkten Dialog zwischen Einrichtungen der Infektionskontrolle sowie der medizinischen Forschung einerseits und den Experten des technischen Datenschutzes und der Kryptografie andererseits. Beide Seiten würden von einem intensiveren Austausch profitieren und könnten vor dem Auftreten einer neuen Krisensituation robuste und bewährte Prozesse etablieren. Sowohl die dadurch mögliche Transparenz und verbesserter öffentlicher Gesundheitsschutz als auch das zeitgleich höhere Schutzniveau der persönlichen Daten würden die in den letzten zwei Jahren bemerkbaren Friktionen reduzieren.

aufgesetzt. (https://www.gesundheitsforschung-bmbf.de/de/smics-smarte-software-gegen-sars-cov-2-11471.php).

Danksagung

Diese Arbeit wurde teilweise aus Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) im Rahmen des Medizininformatik-Initiative Konsortiums HiGHmed, sowie teilweise durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1119 – 236615297 gefördert.

Referenzen

Addanki, Surya, Kevin Garbe, Eli Jaffe, Rafail Ostrovsky, und Antigoni Polychroniadou. 2021. „Prio+: Privacy Preserving Aggregate Statistics via Boolean Shares". 576. https://eprint.iacr.org/2021/576

Agrawal, Shashank, Saikrishna Badrinarayanan, Pratyay Mukherjee, und Peter Rindal. 2020. „Game-Set-MATCH: Using Mobile Devices for Seamless External-Facing Biometric Matching". 1363. https://eprint.iacr.org/2020/1363

Ashur, Tomer, Efrat Cohen, Carmit Hazay, und Avishay Yanai. 2021. „A New Framework for Garbled Circuits". 739. https://eprint.iacr.org/2021/739

Barth-Jones, Daniel. 2012. „The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now". SSRN Scholarly Paper ID 2076397. Rochester, NY: Social Science Research Network.

Braun, Lennart, Daniel Demmler, Thomas Schneider, und Oleksandr Tkachenko. 2020. „MOTION – A Framework for Mixed-Protocol Multi-Party Computation". 1137. https://eprint.iacr.org/2020/1137

Carpov, Sergiu, Nicolas Gama, Mariya Georgieva, und Dimitar Jetchev. 2021. „GenoPPML – a Framework for Genomic Privacy-Preserving Machine Learning". 733. https://eprint.iacr.org/2021/733

Dietrich, Aljoscha, Christian K. Bosse, und Hartmut Schmitt. 2021. „Kontrolle und Überwachung von Beschäftigten". *Datenschutz und Datensicherheit – DuD* 45 (1): 5–10.

Ellinghaus, David, Frauke Degenhardt, Luis Bujanda, Maria Buti, Agustin Albillos, Pietro Invernizzi, Javier Fernandez, u. a. 2020. „The ABO Blood Group Locus and a Chromosome 3 Gene Cluster Associate with SARS-CoV-2 Respiratory Failure in an Italian-Spanish Genome-Wide Association Analysis". *medRxiv*, Juni, 2020.05.31.20114991.

Ellinghaus, David, Frauke Degenhardt, Luis Bujanda, Maria Buti, Agustín Albillos, Pietro Invernizzi, Javier Fernández, u. a. 2020. „Genomewide Association Study of Severe Covid-19 with Respiratory Failure". *New England Journal of Medicine*.

Ellis, Peter James. 2020. „Modelling Suggests Blood Group Incompatibility May Substantially Reduce SARS-CoV-2 Transmission". *medRxiv*, Juli, 2020.07.13.20152637.

European Medicines Agency. 2017. „External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use". European Medicines Agency. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-1.pdf.

Goldreich, O., S. Micali, und A. Wigderson. 1987. „How to Play ANY Mental Game". In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, 218–29. STOC '87. New York, NY, USA: ACM.

Haarbrandt, Birger, Björn Schreiweis, Sabine Rey, Ulrich Sax, Simone Scheithauer, Otto Rienhoff, Petra Knaup-Gregori, u. a. 2018. „HiGHmed – An Open Platform Approach to Enhance Care and Research Across Institutional Boundaries". *Methods of Information in Medicine* 57 (Mai): e66–81.

Hamacher, Kay, Stefan Katzenbeisser, Tobias Kussel, und Sebastian Stammler. 2020. „Genomische Daten und der Datenschutz". *Datenschutz und Datensicherheit – DuD* 44 (2).

Keller, Hannah, Helen Möllering, Thomas Schneider, und Hossein Yalame. 2021. „Balancing Quality and Efficiency in Private Clustering with Affinity Propagation". 825. https://eprint.iacr.org/2021/825

Klar, Manuel, und Jürgen Kühling. 2018. *Datenschutz-Grundverordnung Art. 4 Abs. 1 RN 25*. Herausgegeben von Jürgen Kühling und Benedikt Buchner. C.H. Beck.

Kugelmann, Dieter. 2021. „Ein Jahr Corona-Pandemie – Mehr Digitalisierung braucht effizienteren Datenschutz". *Datenschutz und Datensicherheit – DuD* 45 (5): 297–97.

Landesbeauftragter für den Datenschutz und die Informationsfreiheit Rheinland-Pfalz. o. J. „Corona & Datenschutz". datenschutz.rlp.de. Zugegriffen 10. Mai 2021. https://www.datenschutz.rlp.de/de/themenfelder-themen/corona-datenschutz/

Laufer, Daniel. 2020. „Daten von Infizierten: Polizei sammelt in mehreren Bundesländern Coronavirus-Listen". 2020. https://netzpolitik.org/2020/daten-von-infizierten-polizei-sammelt-in-mehreren-bundeslaendern-coronavirus-listen/

Li, Ninghui, Tiancheng Li, und Suresh Venkatasubramanian. 2007. „T-Closeness: Privacy Beyond k-Anonymity and l-Diversity". In *2007 IEEE 23rd International Conference on Data Engineering*.

Liu, Fei, Yan Jia, und Weihong Han. 2012. „A New K-Anonymity Algorithm towards Multiple Sensitive Attributes". In *2012 IEEE 12th International Conference on Computer and Information Technology*, 768–72.

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, und Muthuramakrishnan Venkitasubramaniam. 2007. „L-Diversity: Privacy beyond k-Anonymity". *ACM Transactions on Knowledge Discovery from Data* 1 (1).

Malkhi, Dahlia, Noam Nisan, Benny Pinkas, und Yaron Sella. 2004. „Fairplay – A Secure Two-Party Computation System". In, 17.

Ohm, Paul. 2010. „Broken promises of privacy: Responding to the surprising failure of anonymization". *UCLA Law Review* 57: 1701.

Oswald, Malcolm. 2013. „Isb1523: Anonymisation Standard for Publishing Health and Social Care Data". National Health Serdice (NHS). https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data

Patra, Arpita, Thomas Schneider, Ajith Suresh, und Hossein Yalame. 2020. „ABY2.0: Improved Mixed-Protocol Secure Two-Party Computation". 1225. https://eprint.iacr.org/2020/1225

Rosulek, Mike, und Lawrence Roy. 2021. „Three Halves Make a Whole? Beating the Half-Gates Lower Bound for Garbled Circuits". 749. https://eprint.iacr.org/2021/749

Sabelfeld, Andrei. 2019. „TOPPool: Time-Aware Optimized Privacy-Preserving Ridesharing". *Proceedings on Privacy Enhancing Technologies* 2019 (4): 93–111.

Schrahe, Dominik, und Thomas Städter. 2021. „COVID-19-Impf- und -Testnachweise". *Datenschutz und Datensicherheit – DuD* 45 (5): 315–19.

Smoltczyk, Maja. 2021. „Datenschutz ist kein Hindernis für digitalen Unterricht – Schulen brauchen Unterstützung". *Datenschutz und Datensicherheit – DuD* 45 (4): 222–22.

Stammler, Sebastian, Tobias Kussel, Phillipp Schoppmann, Florian Stampe, Galina Tremper, Stefan Katzenbeisser, Kay Hamacher, und Martin Lablans. 2020. „Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage Using Secure Multi-Party Computation". *Bioinformatics*.

Sweeney, Latanya. 2000. „Simple Demographics Often Identify People Uniquely". *Health (San Francisco)* 671 (2000): 1–34.

———. 2002. „-Anonymity: A Model for Protecting Privacy". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05): 557–70.

Vogelsang, Lennart, Moritz Lehne, Phillipp Schoppmann, Fabian Prasser, Sylvia Thun, Bj&#246 Scheuermann, rn, und Josef Schepers. 2020. „A Secure Multi-Party Computation Protocol for Time-To-Event Analyses". *Digital Personalized Health and Medicine*, 8–12.

Wu, Yingjie, Xiaowen Ruan, Shangbin Liao, und Xiaodong Wang. 2010. „P-Cover k-Anonymity Model for Protecting Multiple Sensitive Attributes". In *2010 5th International Conference on Computer Science Education*, 179–83.

Yao, Andrew C. 1982. „Protocols for Secure Computations", 5.

———. 1986. „How to Generate and Exchange Secrets". In *27th Annual Symposium on Foundations of Computer Science (Sfcs 1986)*, 162–67.

Zietz, Michael, und Nicholas P. Tatonetti. 2020. „Testing the Association Between Blood Type and COVID-19 Infection, Intubation, and Death". *medRxiv*, Juli.