Dominique Schröder
Lehrstuhl für Angewandte Kryptographie
Fürther Str. 244c, 90429 Nürnberg

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

ChaAC
Chair of Applied Cryptography

Sachverständigengutachten

zum Schutz medizinischer Daten

25. April 2022

Inhalt

1 Zusammenfassung **5**
1.1 Zu meiner Person . . . 6
1.2 Unabhängigkeit . . . 6

2 Re-Personalisierung von Gesundheitsdaten **7**
2.1 Begriffserklärung . . . 7
2.2 Anonymisierung und Pseudonymisierung von Daten . . . 10
2.3 Techniken zur Anonymisierung von Daten . . . 11
2.3.1 Beispiele anonymisierter und pseudonymisierter Datensätze . . . 13
2.4 Angriffe auf anonymisierte Datensätze . . . 15
2.4.1 Der Fall Netflix . . . 15
2.4.2 Der Fall der Datenspende App des RKIs . . . 18
2.5 Zusammenfassung der Ergebnisse . . . 21

3 Alternative Ansätze zur Bereitstellung von (lediglich) pseudonymisierten Datensätzen **23**
3.1 K-Anonymität und verwandte Ansätze . . . 23
3.1.1 K-Anonymität . . . 23
3.1.2 Schwächen von K-Anonymität . . . 24
3.1.3 l-Diversität . . . 24
3.1.4 Weitere Ansätze . . . 26
3.1.5 Allgemeine Probleme bei der Pseudonymisierung und Anonymisierung von Datenbanken . . . 26
3.1.6 Vor- und Nachteile dieser Ansätze . . . 27
3.2 Differential Privacy . . . 27
3.2.1 Differential Privacy Zentralisierter Ansatz . . . 28
3.2.2 Differential Privacy Dezentraler Ansatz . . . 29
3.2.3 Herausforderungen, Vor- und Nachteile dieser Ansätze . . . 30
3.2.4 Die Datenspende App des RKIs: Praktisches Beispiel von Differential Privacy . . . 30
3.3 Berechnung auf verschlüsselten Daten . . . 32
3.3.1 Homomorphe Verschlüsselung . . . 32
3.3.2 Kontrollierte Berechnungen auf Verschlüsselten Daten . . . 35
3.3.3 Vor- und Nachteile dieser Ansätze . . . 37
3.4 Zusammenfassung der Ergebnisse und ein Vergleich der unterschiedlichen Ansätze und deren Qualität in Bezug auf Forschungszwecke . . . 37

4 Sichere Speicherung der Daten **40**
4.1 Entwicklung von Cybersicherheitsangriffen 40
4.1.1 Entwicklung von Cybersicherheitsangriffen 40
4.1.2 Klassifikation der Angriffe 41
4.2 Wert der Daten 42
4.2.1 Finanzieller Schaden eines Cyberangriffs 42
4.2.2 Wert der Daten auf dem Schwarzmarkt 45
4.3 Zentrale vs. Dezentrale Speicherung der Daten 46
4.4 Alternative Ansätze zur Speicherung und Verarbeitung der Daten 48
4.5 Zusammenfassung der Ergebnisse 49

5 Datensammelstelle **50**
5.1 Darstellung des Datentransfers 50
5.2 Intuitive Sicherheitsgedanken für das Design des Datentransfers 51
5.3 Evaluation des Vertrauensmodells des Datentransfers 52
5.4 Alternativer Ansatz für den Datentransfer 53
5.4.1 Alternativer Ansatz: Die Vertrauensstelle als Hüter der Integrität . 53
5.4.2 Zusammenfassung und Vergleich der zwei Ansätze 55

6 Bewertung des Re-Identifikationsrisiko **56**
6.1 Erhobene Daten 56
6.2 Re-Identifikationsrisiko basierend auf soziodemographischen Daten 56
6.3 Re-Identifikationsrisiko basierend auf medizinischen Daten 57
6.3.1 Zusammenfassung der Ergebnisse 59

1 Zusammenfassung

Das vorliegende Sachverständigengutachten zum Schutz medizinischer Daten beschäftigt sich mit der Fragestellung wie gut medizinische Daten anonymisiert werden können, welche (kryptographischen) Techniken zur Bereitstellung dieser Daten genutzt werden können, welche Herausforderungen durch die Speicherung medizinischer Daten entstehen, und ob die Datensammelstelle nach § 303b SGB V technisch erforderlich ist. Zusammenfassend komme ich zu folgenden Ergebnissen:

Anonymisierung / Pseudonymisierung medizinischer Daten. Medizinische Daten spannen einen sehr großen Datenraum auf. Dieser Datenraum umfasst soziodemographische Werte, wie die Größe, Alter und Gewicht, medizinische Informationen über Befunde und Behandlungen, bis hin zur Medikamentierung. Jeder Patient lässt sich als ein Punkt in diesem großen Raum darstellen. Je weiter die Punkte voneinander entfernt sind, umso leichter lässt sich ein Individuum reidentifizieren. Die Ergebnisse aus Kapitel 2 zeigen, dass für die Reidentifizierung eines Individuums nur sehr wenige Datenpunkte benötigt werden. Dies wird anhand von zwei Beispielen verdeutlicht: die Reidentizierungsangriffe gegen anonymisierte Datensätze von Netflix und der Datenspende App des Robert-Koch-Institutes. Aufgrund der Feingranularität der medizinischen Daten komme ich zu dem Schluss, dass die Re-Identifikation ohne großen Aufwand möglich ist, sofern nicht weitere Maßnahmen zum Schutz der Privatsphäre ergriffen werden (siehe Abschnitt 3.2). Es muss davon ausgegangen werden, dass das „schlüsselabhängige Verfahren zur Pseudonymisierung“, welches keine Rückschlüsse „auf das Lieferpseudonym oder die Identität des Versicherten“ erlaubt, wie es § 303c Abs. 2 SGB V vorschreibt, keinen Schutz bietet, da die Re-Identifikation über die anderen Merkmale in den Daten erfolgt.

Alternative Ansätze. Kapitel 3 stellt verschiedene alternative Ansätze zur Anonymisierung und Pseudonymisierung vor und vergleicht diese hinsichtlich der Anwendbarkeit, Sicherheitsgarantien und Qualität. Die Ergebnisse aus Kapitel 3 zeigen, dass es eine Vielzahl von kryptographischen Techniken zur Anonymisierung und Pseudonymisierung von medizinischen Daten gibt. Des Weiteren werden Ansätze zur sicheren Berechnung auf verschlüsselten Daten beschrieben und deren Vor- und Nachteile besprochen. Die Ergebnisse dieses Kapitels zeigen deutlich, dass in vielen Punkten der Stand des Wissens nicht befolgt wird.

Sichere Speicherung. Im Kapitel 4 werden die Vor- und Nachteile einer zentralen Speicherung medizinischer Daten diskutiert und mit dezentralen Ansätzen verglichen. Insbesondere wird auch auf die Frage eingegangen, ob die zentrale Speicherung der Daten zur Zusammenführung notwendig ist. Die Ergebnisse dieses Abschnittes

zeigen, dass die Risiken der zentralen Speicherung der Daten nicht vertretbar sind, da die gleiche Funktionalität auch dezentral realisiert werden kann.

Notwendigkeit der Datensammelstelle. Auf die Frage nach der Notwendigkeit einer zentralen Datensammelstelle wird in Kapitel 5 eingegangen. Die Ergebnisse dieses Abschnittes zeigen, dass eine zentralisierte Zusammenführung der Daten nicht notwendig ist und Konsistenz- und Plausibilitätsprüfungen auch dezentral durchgeführt werden können. Des Weiteren komme ich in diesem Abschnitt zu dem Schluss, dass der Entwurf grundlegende Prinzipien der IT Sicherheit ignoriert und ohne die Entwicklung eines formalen Sicherheitsmodells erfolgt, wodurch die Schutzziele nicht nachvollziehbar sind. Dies ist vergleichbar mit dem Bau eines Hauses bei dem weder eine Berechnung der Statik, noch die Erstellung eines Brandschutzkonzeptes vor dem Bau durchgeführt wurde. Dies entspricht nicht dem aktuellen Stand des Wissens und ignoriert die Forschungserkenntnisse der letzten 20-30 Jahre im Bereich der IT Sicherheit.

Bewertung des Re-Identikationsrisiko Im Kapitel 6 wird auf Schwierigkeit der Bewertung des Re-Identifikationsrisiko durch das Forschungsdatenzentrum eingegangen, dabei komme ich zum Schluss, dass das Forschungsdatenzentrum das Re-Identifikationsrisiko höchstwahrscheinlich nicht bewerten kann. Diese Bewertung begründet sich in der Tatsache, dass sehr wenige Datenpunkte zur Re-Identifikation ausreichen, es an Forschungsergebnissen in diesem Gebiet mangelt, das Hintergrundwissen des Angreifers nicht abgeschätzt werden kann und es insbesondere im Bereich der Medizin an Erfahrung mangelt, welche Daten zur Re-Identifikation benutzt werden können.

1.1 Zu meiner Person

Herr Dominique Schröder leitet den Lehrstuhl für Angewandte Kryptographie an der Friedrich-Alexander-Universität Erlangen-Nürnberg seit 2016. Er war als Einzelsachverständiger im Gesundheitsausschuss des Deutschen Bundestages zum Entwurf des DVGs geladen. In seiner Forschung beschäftigt er sich mit der Entwicklung von Techniken zum Schutz der Privatsphäre. Die Ergebnisse seiner Forschung wurden durch zahlreiche Preise ausgezeichnet, wie dem Feodor-Lynnen Forschungsstipendum der Humboldt-Gesellschaft oder dem Intel Early Career Faculty Award.

Die Ansichten in diesem Gutachten spiegeln die persönliche Meinung des Autors wider und sind nicht zwingend die Ansichten der Friedrich-Alexander Universität Erlangen-Nürnberg.

1.2 Unabhängigkeit

Im Zuge dieses Gutachtens wurden mir lediglich die Fragen vorgegeben. Auf die Beantwortung der Fragen und damit auf die Ergebnisse dieses Gutachtens wurde durch Dritte kein Einfluss genommen.

2 Re-Personalisierung von Gesundheitsdaten

Fragestellung: Wie leicht lassen sich pseudonymisierte (Gesundheits-)Daten re-personalisieren?

Zur Beantwortung dieser Frage gliedert sich dieses Kapitel in folgende Abschnitte. Im Abschnitt 2.1 werden grundlegende Begriffe eingeführt und voneinander abgegrenzt. Dies dient nicht nur dem Verständnis dieses Kapitels, sondern trägt auch zum Verständnis des gesamten Gutachtens bei. Die allgemeine Schwierigkeit Daten zu anonymisieren wird in Abschnitt 2.2 thematisiert, gefolgt von den Techniken zur Anonymisierung in Abschnitt 2.3. Die Grenzen dieser Techniken werden anhand von praktischen Beispielen in Abschnitt 2.4 aufgezeigt, im Anschluss erfolgt die Zusammenfassung der Ergebnisse in Abschnitt 2.5.

2.1 Begriffserklärung

In diesem Abschnitt werden die grundlegenden Begriffe *Anonymisierung*, *Pseudonymisierung*, *De-anonymisierung* und *Verschlüsselung* beschrieben und definiert. Diese Beschreibung hat zum Ziel, ein gemeinsames Verständnis für die unterschiedlichen Begriffe zu schaffen und eine präzise Abgrenzung zu erhalten. Die folgenden Definitionen der Begriffe sind die eines Kryptographen. Nach Bretthauer und Spiecker gen. Döhmann bedarf es noch einer (juristischen) Klärung dieser grundsätzlichen Begriffe im Kontext der Gesundheitsdatenverarbeitung [1].

Anonymisierung

Zur Definition des Begriffes der Anonymisierung muss zunächst die Bedeutung des Begriffes der *Anonymität* geklärt werden. Der Begriff der Anonymität wird sowohl für Personen als auch für Daten verwendet. Intuitiv versteht man unter der Anonymität von Personen, dass diese nicht identifiziert werden können. Dies bedeutet, dass die Zuordnung einer Sache oder einer Information zu einer bestimmten Person mit verhältnismäßig großem Aufwand nicht möglich sein soll.* Der Aufwand bezieht sich dabei auf die Zeit, Kosten und Arbeitskraft oder die Rechenleistung. Kryptographen und Juristen scheinen eine andere Auffassung zu haben, was unter dem Begriff des „verhältnismäßig großem

*Die DS-GVO legt den Begriff der Anonymität sehr strikt aus indem die bloße Existenz einer Zuordnung ausgeschlossen wird [2].

Aufwand“ verstanden wird: Kryptographen gehen immer vom schlechtesten Fall aus und fassen Parameter wie Zeit, Kosten und mögliches Hintergrundwissen in dem Begriff von Rechenoperationen zusammen. Das heißt, Kryptographen gehen davon aus, dass der Angreifer über beliebig finanzielle Ressourcen und über ein beliebig großes Hintergrundwissen verfügt, jedoch nur eine bestimmte Anzahl von Rechenoperationen durchführen kann, um diese Informationen zu verarbeiten. Die Anzahl der zulässigen Rechenoperationen steigt mit der Entwicklung der Hardware und wird als Sicherheitsparameter definiert. Aktuell würde man zum Beispiel von einem Sicherheitsparameter von 2^{128} ausgehen[†]. Dies bedeutet, dass der Angreifer unter Einbindung von *beliebigen* Informationen und Ressourcen seinen Angriff in höchstens 2^{128} Rechenoperationen durchführen darf. Ein System gilt als sicher, wenn der Angreifer nach diesen Schritten keine nichttriviale Information *hinzu gewonnen* hat. Handelt es sich bei dem Angriff zum Beispiel um einen Deanonymisierungsangriff auf einen Patienten und steht zum Beispiel in dem Hintergrundwissen der Names des Patienten drin, dann kann der Angreifer nicht mehr erfolgreich sein, da er dieses Wissen bereits (trivial) über das Hintergrundwissen erhalten hat.

Juristisch scheint der Begriff des „verhältnismäßig großem Aufwands“ deutlich differenzierter zu sein, indem eine „umfassende Betrachtung verschiedener Aufwandsfaktoren“ erstellt wird, auf dessen Basis soll dann eine eine Wahrscheinlichkeitsprognose angefertigt werden, die zu einer Kosten-Nutzen-Analyse führt [3].

Damit eine Person nicht identifiziert werden kann, wird immer eine Menge von Personen benötigt in welcher die Person „verschwindet“. Diese Menge wird als Anonymitätsmenge (engl. *anonymity set*) bezeichnet. Besteht die Menge aus lediglich einer Person, so kann diese nicht anonym sein. In diesem Gutachten verwende ich den Begriff der Anonymität im kryptographischen Sinne, d.h., ein effizienter Algorithmus soll nicht in der Lage sein, Daten zu einer Person mit hoher Wahrscheinlichkeit zuzuordnen.

Pseudonymisierung

Damit Daten einem Individuum zugeordnet werden können, müssen die Daten über bestimmte Merkmale verfügen, die eine Zuordnung ermöglichen. Dabei wird im Allgemeinen zwischen drei Arten unterschieden: *Identifikatoren, Quasi-Identifikatoren* und *sensible Attribute*. Identifikatoren sind Merkmale, die eine direkte Zuordnung erlauben, wie zum Beispiel die Nummer des Personalausweises oder der Fingerabdruck.

Quasi-Identifikatoren sind Merkmale, die die Menge der möglichen Personen stark einschränken, sodass eine Zuordnung durch eine Kombination von mehreren Quasi-Identifikatoren wahrscheinlich ist. Zu den Quasi-Identifikatoren zählt zum Beispiel das Geschlecht, die Postleitzahl, der Geburtstag, der Beruf usw. Ist zum Beispiel nur das Geschlecht bekannt, so reduziert sich die Menge der möglichen Individuen um ca. 50% (unter der Annahme, dass die Verteilung von Männern und Frauen ungefähr gleich ist [4]). Ist nicht nur das Geschlecht bekannt, sondern auch das Alter und die Postleitzahl, so lässt sich die Menge an möglichen Individuen stark einschränken [5]. Sensible Attribute

[†]Die Anzahl von 2^{128} Operationen entsprechen 340.282.366.920.938.463.463.374.607.431.768.211.456 einzelne Rechenschritte.

speichern persönliche und schützenswerte Informationen, wie zum Beispiel Krankheiten.

Der Prozess der Pseudonymisierung beschreibt nun, dass Identifikationsmerkmale durch zufällige Elemente ersetzt werden. Laut Art. 4 Nr. 5 DSGVO muss dieser Prozess sicherstellen, dass „personenbezogene Daten ohne Hinzuziehen zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können [...]“.

Abgrenzung Anonyme und Pseudonyme Daten

Auf den ersten Blick erscheint die Abgrenzung zwischen anonymen und pseudonymen Daten schwierig. Bei pseudonymisierten Daten wurden Identifikatoren durch zufällige Werte ersetzt, sodass die Identifikation eines Individuums für denjenigen, der die Daten verarbeitet, mit vertretbaren Mitteln nicht möglich sein soll. Dies bedeutet insbesondere, dass die Identifikation eines Individuums durch Dritte, die potenziell andere Daten über das Individuum haben, durchaus möglich sein könnte. Der Begriff von anonymen Daten ist deutlich stärker, da die Möglichkeit der Re-Identifikation, unabhängig von den Ressourcen des Angreifers wie zum Beispiel der Zeit, der Rechenleistung und dem Zusatzwissen über das Individuum, generell nicht möglich sein darf. Daraus folgt, dass anonyme Daten auch immer pseudonyme Daten sind, jedoch sind pseudonyme Daten nicht zwingend anonym.

De-anonymisierung

Das Ziel der De-Anonymisierung/Re-Identifikation ist die möglichst eindeutige Zuordnung von anonymisierten oder pseudonymisierten Daten zu einem bestimmten Individuum. Um anonymisierte bzw. pseudonymisierte Daten zu de-anonymisieren, werden die Daten in der Regel mit weiteren Informationen verknüpft. Das Ziel dieser Verknüpfung folgt einer Art Ausschlussprinzip, je kleiner die Anonymitätsmenge ist, also die Menge der Personen, die infrage kommen, umso genauer kann man eine bestimmte Person identifizieren.

Verschlüsselung

Der Klartext bezeichnet einen Text in seiner ursprünglichen Form. Unter einer Verschlüsselung wird eine kryptographische Operation bezeichnet, die mittels kryptographischen Schlüsseln einen Klartext in einen sogenannten Chiffretext überführt. Der Chiffretext gibt keinerlei Informationen über den Klartext preis (bis auf Informationen, die sich trivial aus dem Chiffretext ableiten lassen, wie zum Beispiel die Länge des Klartextes). Die Sicherheit beruht lediglich auf der Geheimhaltung der kryptographischen Schlüssel. Dies bedeutet, dass das verwendete Verschlüsselungsverfahren und alle anderen Parameter, wie zum Beispiel der Sicherheitsparameter, die Schlüssellänge, der Modus in dem das Verschlüsselungsverfahren betrieben wird, oder die verwendete Hardware, bekannt sind.

2.2 Anonymisierung und Pseudonymisierung von Daten

Wie im Abschnitt 2.1 beschrieben, ist das Ziel der Anonymisierung von personenbezogenen Daten in eine Form zu überführen, die unabhängig von den Ressourcen und Zusatzwissen keine Zuordnung von Daten zu einem Individuum erlaubt. Hingegen ist bei der Pseudonymisierung von Daten eine Zuordnung prinzipiell möglich, jedoch soll diese in dem vorgesehenen Verwendungsbereich nur mit einem unverhältnismäßig hohem Aufwand möglich sein. Dies bedeutet insbesondere, dass die Re-Identifikation der Daten durch Dritte durchaus leicht sein kann.

Die Anonymisierung von Daten ist aus folgenden Gründen im Allgemeinen sehr schwierig:

- Die Informationstheorie beschreibt ein statistisches Maß für die Information, die in einer Nachricht enthalten ist. Es gibt ein natürliches Spannungsfeld zwischen Anonymität und dem Informationsgehalt. Eine perfekte Anonymität kann nur dann erreicht werden, wenn die Daten keine Informationen mehr beinhalten. Wenn die Daten jedoch keinerlei Informationen beinhalten, dann sind diese Daten auch wertlos. Um aus der Analyse der Daten einen Mehrwert zu generieren, muss folglich zwingend ein gewisser Abstrich in Bezug auf die Anonymität hingenommen werden.

- Die interessanten Aspekte an den Daten sind oftmals die sensiblen Attribute oder Merkmale, die als Quasi-Identifikatoren dienen. Würden alle sensiblen Attribute und Quasi-Identifikatoren aus den Daten entfernt werden, so blieben am Ende keine oder nur sehr wenig Informationen übrig.

- Um ein Individuum eindeutig zu identifizieren, sind nur sehr wenige Datenpunkte notwendig. Dies wird im Unterabschnitt 2.4.2 anhand unserer Sicherheitsanalyse der „Datenspende App“ des Robert-Koch-Institutes verdeutlicht.

Die oben genannten Aspekte gelten für beliebige Daten und beziehen nicht die Besonderheit medizinischer Daten mit ein. Sowohl die Anonymisierung als auch die *zuverlässige* Pseudonymisierung von medizinischen Daten ist aus meiner Sicht ungleich schwieriger, da folgende Aspekte einbezogen werden müssen:

- In vielen Fällen wissen wir nicht, welche Merkmale als Quasi-Identifikatoren dienen könnten. Dies ist insofern ein Problem, da die Merkmale im Zuge der Anonymisierung (bzw. Pseudonymisierung) übersehen wurden. Dies ist aus meiner Sicht ein offenes Problem für die Wissenschaft, welches zwingend mehr Forschung benötigt. Da die Quasi-Identifikatoren nicht bekannt sind, wird oftmals auch keine Pseudonymisierung durchgeführt und folglich können aus publizierten Daten, Rückschlüsse auf Individuen gemacht werden.

- Medizinische Daten können nicht nur Aussagen über ein Individuum treffen, sondern können auch Informationen über Dritte preisgeben. Am einfachsten sieht man diese Tatsache an genetischen Daten. Da das Genom ungefähr zu gleichen Teilen

aus dem Genom der Mutter und des Vaters besteht, gibt die Veröffentlichung gleichzeitig Informationen über die (Groß-)Eltern, Kinder und Enkel usw. preis. Andere Beispiele sind Krankheiten die genetisch bedingt sind oder familiär auftreten. Zu diesen zählt zum Beispiel die Krankheit Demenz. Ist der Grund der Krankheit genetisch bedingt, „sind bei rund 30 Prozent aller Fälle bei einer auftretenden Häufung auch enge Familienmitglieder betroffen“[6]. Ein weiteres Beispiel ist erblich bedingter Brustkrebs, „bei etwa einem Viertel aller Frauen mit Brustkrebs treten vermehrt Brustkrebsfälle in der Familie auf“ [7].

Im Folgenden werden gängige Techniken aus der Praxis zur Anonymisierung und Pseudonymisierung von Daten vorgestellt.

2.3 Techniken zur Anonymisierung von Daten

Die Techniken zur Anonymisierung von Daten hängen oftmals von dem Kontext ab und werden auf die Bedürfnisse der jeweiligen Anwendung zugeschnitten [2]. In vielen Fällen gibt es auch keine Standards, sodass eine allumfassende Klassifizierung dieser Techniken nicht möglich ist. Vergleicht man jedoch die gängigen Ansätze, so lassen sich diese in fünf Kategorien unterteilen:

Informationsunterdrückung Bei dem Ansatz der Informationsunterdrückung werden sensible Attribute aus dem Datensatz gelöscht.

Auf der einen Seite, aus Sicht der Privatsphäre, hat dies den Vorteil, dass keine Rückschlüsse auf eine Person *aufgrund des gelöschten Attributes* mehr möglich sind. Auf der anderen Seite gehen damit alle Informationen des Attributes verloren und damit sinkt die Güte der Daten. Dies kann insbesondere im medizinischen Bereich schwerwiegende Auswirkungen auf die Diagnose haben und unter Umständen zu einer falschen Diagnose führen. Betrachten wir zum Beispiel eine Menge von Patienten von denen 99% unter erblich bedingter Demenz leiden. Damit die Teilnehmer, die nicht unter erblich bedingter Demenz leiden, nicht de-anonymisiert werden können, wird das Attribut „erblich bedingte Demenzerkrankung“ entfernt. Nun wird auf diesem Datensatz zufällig eine Studie über Demenzerkrankungen durchgeführt. Diese Studie kommt zu dem Schluss, dass 99% der Menschen in der Bevölkerung unter Demenz leiden, da dies das Ergebnis der Studie ergeben hat. Jedoch ist dieses Ergebnis offensichtlich falsch, da die Stichprobe, also die Menge der Patienten, ungünstig gewählt wurde. Aufgrund der Informationsunterdrückung des Attributes „erblich bedingte Demenzerkrankung“, konnte dies jedoch nicht festgestellt werden. Dieses Beispiel zeigt, dass diese Technik im medizinisches Bereich oft nicht eingesetzt werden kann.

Generalisierung Das Ziel der Generalisierung eines Attributes besteht in der Vergrößerung der möglichen Anonymitätsmenge in Bezug auf das Attribut. Speichert zum Beispiel ein Datenbankeintrag das genaue Geburtsdatum einer Person, so könnte dieses im ersten Schritt auf das Alter generalisiert werden. Dadurch erhöht sich die

Menge an möglichen Personen von denen, die an einem bestimmten Tag Geburtstag haben, auf die Menge an Personen, die in einem bestimmten Jahr Geburtstag haben. Eine weitere Generalisierung könnte zum Beispiel die Angabe einer Spanne sein, in der die Personen Geburtstag haben, beispielsweise alle Menschen im Alter von 25 - 30 Jahren.

Die Technik der Generalisierung reduziert natürlich die Güte der Daten, jedoch erlaubt dieses Verfahren oftmals statistische Aussagen über eine große Menge zu treffen.

Verrauschung Unter dem Begriff der Verrauschung wird das gezielte Verrauschen von Attributen bezeichnet. Bei einer Verrauschung der Daten werden gezielt (kleine) Fehler in die Attribute hinzugefügt. Ist eine Person zum Beispiel 25 Jahre alt, so könnte das Alter mit 23 oder 28 Jahren angegeben werden.

Auf den ersten Blick mag die Technik der Verrauschung nicht sinnvoll erscheinen. Durch diese Technik werden Attribute gezielt verfälscht und damit droht die Gefahr, dass es bei Auswertungen von medizinischen Daten zu falschen Aussagen kommt. Die Verrauschung von Daten ist jedoch ein wesentlicher Bestandteil der *Differential Privacy*, welche im Abschnitt 3.2 genauer beschrieben wird und sich als sehr effektive Technik für statistische Auswertungen (medizinischer) Daten erweist.

Anatomisierung Die Technik der Anatomisierung besteht im Wesentlichen aus zwei Schritten [8]: Im ersten Schritt wird eine Tabelle gezielt aufgeteilt, sodass die eine Tabelle die Werte der Quasi-Identifikatoren und die zweite Tabelle die sensitiven Attribute speichert. Die einzelnen Tabellen werden nun mit einer Gruppen-Identifikationsnummer miteinander verknüpft. Diese Verknüpfung führt dazu, dass sensitive Daten zwar einer Gruppe zugeordnet werden können, jedoch geht die eindeutige Zuordnung zu einem Individuum verloren.

Der Vorteil dieses Ansatzes besteht darin, dass keine Fehler eingeführt werden und kein Element weggelassen wird. In Bezug auf die Anonymität wird ein schwächeres Niveau erreicht, da mit einer gewissen Wahrscheinlichkeit Rückschlüsse auf Individuen getroffen werden können.

Permutation Eine weitere Technik zur Anonymisierung von Daten ist die Technik der Permutation, die ebenfalls aus zwei Schritten besteht. Im ersten Schritt werden die Elemente in Klassen eingeteilt, zum Beispiel könnte eine Klasse aus allen Einträgen von Frauen und die zweite Klasse aus allen Einträgen von Männern bestehen. Im zweiten Schritt werden dann die Elemente, die als Quasi-Identifikatoren bestimmt wurden, permutiert, also zufällig gemischt.

Die grundlegende Überlegung hinter dieser Technik ist die Folgende: Die Einteilung in Gruppen teilt die Menge in Untergruppe, die man trivial unterscheiden kann, wie zum Beispiel die Gruppe der Frauen und Männer. Diese Gruppen stellen also die höchste erreichbare Anonymitätsmenge dar. Innerhalb dieser Anonymitätsmenge wird im zweiten Schritt die Verknüpfung von einer Person mit einem

Quasi-Identifikator aufgehoben indem die Quasi-Identifikatoren permutiert werden. Dieses Prinzip kann an einem einfachen Beispiel verdeutlicht werden bei dem der Quasi-Identifikator angibt, ob der Patient Brustkrebs hat oder nicht. Durch die Permutation bleibt die Menge an Patienten mit Brustkrebs gleich, jedoch kann man nicht sagen, ob der Eintrag wirklich zu dem Patient gehört. Dieses Beispiel verdeutlicht auch warum erst nach der Aufteilung permutiert wird und nicht vorher. Würde vor der Aufteilung permutiert, dann könnten Widersprüche entstehen, da nun auch männliche Patienten vermehrt Brustkrebs haben. Zwar ist dies möglich, jedoch statistisch gesehen unwahrscheinlich. Folglich dient die Kombination beider Schritte dazu, dass eine maximale Anonymitätsmenge bestimmt wird (Schritt 1) und eine Konsistenz der Gesamtaussage in dieser Menge erhalten bleibt (Schritt 2).

Die Vor- und Nachteile dieser Technik sind vergleichbar mit denen der Anatomisierung [8].

Basierend auf diesen Grundtechniken wurden weitere Verfahren entwickelt, wie zum Beispiel das Verfahren der *Angelization* oder des *Slicing*, die als eine Kombination der oben genannten Techniken angesehen werden können oder vielmehr sich daraus ableiten lassen [9, 10]. Beispiele für die einzelnen Techniken können dem Unterabschnitt 2.3.1 entnommen werden.

2.3.1 Beispiele anonymisierter und pseudonymisierter Datensätze

In diesem Abschnitt werden die grundlegenden Begriffe und Techniken vorgestellt deren Ziel die Anonymisierung von Daten ist. Diese werden anhand von einfachen Beispielen verdeutlicht. Zunächst betrachten wir die Datenbank in Tabelle 2.1.

Vorname	Nachname	Geburtsdatum	Geschlecht	Adresse	PLZ	Ort	Diagnose
Alex	Schmidt	11.01.1970	M	Grünstraße 12	38102	Braunschweig	Grippe
Maria	Francis	02.12.1980	W	Schlossplatz 1	91054	Erlangen	COVID-19
Carol	Schubert	01.04.1971	W	Einsteinstraße 130	91074	Herzogenaurach	Krebs
Jürgen	Schulz	11.08.1952	M	Kaiserhof	23746	Kellenhusen	COVID-19
Marie	Meier	13.03.1967	W	Berliner Ring 2	38436	Wolfsburg	Krebs
Julie	Smith	19.04.2000	W	Konrad-Zuse-Straße 3	91052	Erlangen	Krebs
Jon	Miller	02.09.1945	M	Grace-Hopper-Straße	23562	Lübeck	COVID-19

Tab. 2.1: Einfache medizinische Datenbank ohne Modifikationen.

In den beiden ersten Spalten steht der Name des Patienten. Der vollständige Name wird oftmals als direkter Identifikator angesehen, obwohl eine eindeutige Zuordnung nicht zwingend gegeben ist (im Gegensatz zu der Nummer des Personalausweises, welche zwingend eindeutig ist). Zu den Quasi-Identifikatoren zählen der Geburtstag, das Geschlecht, die Adresse, die Postleitzahl und der Ort. Die letzte Spalte gibt mit der Diagnose ein sensibles Attribut an. Im ersten Schritt werden durch das Prinzip der Informationsunterdrückung die Identifikatoren entfernt beziehungsweise durch eine zufällige Zahl ersetzt. Die resultierende Tabelle ist in Tabelle 2.2 dargestellt.

ID	Geburtsdatum	Geschlecht	Adresse	PLZ	Ort	Diagnose
00001	11.01.1970	M	Grünstraße 12	38102	Braunschweig	Grippe
00002	02.12.1980	W	Schlossplatz 1	91054	Erlangen	COVID-19
00003	01.04.1971	W	Einsteinstraße 130	91074	Herzogenaurach	Krebs
00004	11.08.1952	M	Kaiserhof	23746	Kellenhusen	COVID-19
00005	13.03.1967	W	Berliner Ring 2	38436	Wolfsburg	Krebs
00006	19.04.2000	W	Konrad-Zuse-Straße 3	91052	Erlangen	Krebs
00007	02.09.1945	M	Grace-Hopper-Straße	23562	Lübeck	COVID-19

Tab. 2.2: Im ersten Schritt werden die Namen durch Identifikationsnummer ersetzt.

Im zweiten Schritt wird das Prinzip der Generalisierung angewendet. Anstatt das Geburtsdatum genau anzugeben, wird dieses durch eine Spanne ersetzt. Des Weiteren werden die Straße und die letzten drei Ziffern der Postleitzahl entfernt. Das Ergebnis dieses Prozesses ist in Tabelle 2.3 dargestellt.

ID	Alter	Geschlecht	PLZ	Diagnose
00001	[50-59]	M	38***	Grippe
00002	[40-49]	W	91***	COVID-19
00003	[40-49]	W	91***	Krebs
00004	[70-79]	M	23***	COVID-19
00005	[50-59]	W	38***	Krebs
00006	[40-49]	W	91***	Krebs
00007	[70-79]	M	23***	COVID-19

Tab. 2.3: Diese Tabelle zeigt das Ergebnis nach den ersten beiden Operationen.

Im nächsten Schritt soll nun das Prinzip der Permutation verdeutlich werden, dazu werden die Daten in Tabelle 2.3 im ersten Schritt in zwei Klassen aufgeteilt: die der weiblichen und der männlichen Patienten. Um die Lesbarkeit zu erhöhen wurden beide Klassen farblich hervorgehoben (siehe Tabelle 2.4a). Im zweiten Schritt erfolgt die Permutation der Quasi-Identifikatoren. In dem Beispiel wird das Alter der ID 0001 mit dem der ID 0004 getauscht, gleiches git für die IDs 0003 und 0005. Des Weiteren wird die PLZ der IDs 0005 und 0006 getauscht. Das Ergebnis dieser Operationen ist in Tabelle 2.4b dargestellt.

Im Folgenden wird nun das Prinzip der Anatomisierung dargestellt, dabei wird erneut von den nicht permutierten Daten ausgegangen (siehe Tabelle 2.3). Dazu werden die Einträge zunächst in Gruppen eingeteilt, beispielsweise in die Gruppe der weiblichen und männlichen Patienten. Jede dieser Gruppen erhält eine Gruppen-Identifikationsnummer (GID), einfachheitshalber ordnen wir der Gruppe der weiblichen Patienten die Identifikationsnummer G1 zu und den männlichen Patienten G2. Im nächsten Schritt wird die Tabelle aufgeteilt, sodass die zweite Tabelle die sensitiven Attribute (Diagnose) und deren Anzahl speichert. Das Ergebnis dieser Operation ist im rechten Teil der Tabelle 2.5

ID	Alter	Geschlecht	PLZ	Diagnose
00001	[50-59]	M	38***	Grippe
00002	[40-49]	W	91***	COVID-19
00003	[40-49]	W	91***	Krebs
00004	[70-79]	M	23***	COVID-19
00005	[50-59]	W	38***	Krebs
00006	[40-49]	W	91***	Krebs
00007	[70-79]	M	23***	COVID-19

(a) Einteilung der Patienten in zwei Gruppen.

ID	Alter	Geschlecht	PLZ	Diagnose
00001	[70-79]	M	38***	Grippe
00002	[40-49]	W	91***	COVID-19
00003	[50-59]	W	91***	Krebs
00004	[50-59]	M	23***	COVID-19
00005	[40-49]	W	91***	Krebs
00006	[40-49]	W	38***	Krebs
00007	[70-79]	M	23***	COVID-19

(b) Permutation der Quasi-Identifikatoren innerhalb der Klassen.

Tab. 2.4: Darstellung des Prinzips der Permutation.

dargestellt.

ID	Alter	Geschlecht	PLZ	GID
00001	[50-59]	M	38***	G2
00002	[40-49]	W	91***	G1
00003	[40-49]	W	91***	G1
00004	[70-79]	M	23***	G2
00005	[50-59]	W	38***	G1
00006	[40-49]	W	91***	G1
00007	[70-79]	M	23***	G2

GID	Diagnose	Anzahl
G2	Grippe	1
G1	Krebs	3
G1	COVID-19	1
G2	COVID-19	2

Tab. 2.5: In diesem Schritt wir das Prinzip der Anatomisierung angewendet.

2.4 Angriffe auf anonymisierte Datensätze

Auf den ersten Blick erscheinen die Techniken zur Anonymisierung von (medizinischen) Daten sinnvoll und effektiv. In diesem Abschnitt stellen wir verschiedene Angriffe auf vermeintlich anonymisierte Daten vor. Der erste Fall ist die sogenannte *Netflix Challenge*, bei welcher Netflix einen Teil seiner Nutzerdaten anonymisiert veröffentlicht hatte mit dem Ziel, einen besseren Algorithmus zum Vorschlagen neuer Filme zu finden. In dem zweiten Beispiel gehen wir auf die Datenspende App des Robert-Koch-Institutes ein. Dieser Abschnitt umfasst aktuelle Forschungsergebnisse, die unter meiner Verantwortung entwickelt wurden.

2.4.1 Der Fall Netflix

Netflix ist ein amerikanisches Medienunternehmen, welches sich auf das Streaming von Videofilmen spezialisiert hat. Im Oktober 2006 startete die sogenannte Netflix Challenge (oder auch Netflix Preis). In einem öffentlichen Wettbewerb versprach Netflix dem Gewinner eine Million USD für einen Algorithmus, der dem Benutzer bessere Filmvorschläge

basierend auf dem individuellen Interesse macht. Zu diesem Zweck veröffentlichte Netflix vermeintlich anonymisierte Daten seiner Nutzer. Insgesamt stellte Netflix 100.480.507 Datenbankeinträge mit Filmbewertungen von 480.189 Benutzern über 17.770 Filme zur Verfügung. Zum besseren Schutz der Privatsphäre wurden die Daten leicht verrauscht, durch das Entfernen von Bewertungen, Einfügen von anderen Bewertungen und anderen Bewertungstagen. Bei diesen Daten handelte es sich um die echten Nutzerdaten, die

Benutzer	Film	Datum der Bewertung	Bewertung
2532865	4500	2005-07-26	5
573364	4500	2005-06-20	3
1696725	4500	2004-02-27	3
1253431	4500	2004-03-31	3
1265574	4500	2003-09-01	2
1049643	4500	2003-11-15	1

Tab. 2.6: Beispieldaten aus dem Netflixpreis [11].

Netflix im Zeitraum von Oktober 1998 und Dezember 2005 gesammelt hat. Diese Daten bestehen aus einer Identifikationsnummer des Kunden, einer ID des Filmes (in diesem Fall ist es der Film mit der ID 4500, „Les Dames du Bois de Boulogne“), das Datum der Bewertung und einer Bewertung von 1 bis 5 Sternen. Zum Schutz der Privatsphäre der Kunden handelte es sich nicht um die echte Identifikationsnummer des Kunden, sondern um eine zufällig gewählte.

De-anonymisierungs-Angriffe

Auf den ersten Blick erscheinen die Daten vollkommen harmlos und die De-anonymisierung eines Kunden, das heißt, die Zuordnung der Vorlieben eines Kunden für bestimmte Filme zu einer realen Person, erscheint nicht möglich. Im Jahre 2008 zeigten Narayanan und Shmantikov, dass diese Intuition falsch ist, indem sie einen fehlertoleranten Algorithmus entwickelten, der die Daten mit den Bewertungen aus der öffentlichen Filmbewertungsdatenbank „Internet Movie Database“(IMDB) abgleicht [12]. Ein Algorithmus ist fehlertolerant, wenn dieser teilweise fehlerhafte Eingaben erhält, aber dennoch das richtige Ergebnis berechnet. Die Autoren zeigten in ihrer Arbeit, dass mit lediglich 8 Bewertungen (von denen 2 falsch sein können) und bei denen das Datum der Bewertung bis zu 14 Tage vom eigentlichen Datum abweicht, 99% der Datenbankeinträge eindeutig zugeordnet werden können [12]. Selbst wenn lediglich 2 Bewertungen abgegeben wurden (mit einer Fehlertoleranz im Datum von zwei Tagen), können immer noch 68% der Datenbankeinträge eindeutig zugeordnet werden [12]. Sollte kein Datum der Bewertung bekannt sein, so kann ein Kunde immer noch gut de-anonymisiert werden, wenn bekannt ist, dass der Kunde Filme außerhalb der Top 500 bewertet hat.

Interpretation der Ergebnisse und Implikationen für Medizinische Daten

Die Ergebnisse von Narayanan und Shmantikov erlauben keinen direkten Rückschluss auf eine natürliche Person, das war auch nicht das Ziel. Vielmehr ist die Kernaussage

der Forschungsarbeit und dessen Implikation von Bedeutung:

- Der Ausgangspunk der Arbeit ist die Veröffentlichung einer sehr großen vermeintlich anonymisierten Datenbank.
- Die Forscher stellen die Frage, ob diese große anonymisierte Datenbank mit einer kleinen so verknüpft werden kann, dass eine eindeutige Zuordnung entsteht.
- Die Arbeit zeigt, dass sehr wenig Datenpunkte mit einer geringen Variabilität, von denen auch einige fehlerhaft sein können, in einer sehr großen Menge zur De-Anonymisierung einer Person ausreichen. Geringe Variabilität bezieht sich zum Beispiel auf das Spektrum der möglichen Bewertungen zwischen 1 und 5. Im Falle von Medikamenten und deren Kombination wird sicher eine größere Variabilität erreicht, was eine Re-Identifikation vereinfacht.
- Des Weiteren verdeutlicht diese Arbeit, dass jedes Datum, unabhängig von dem Inhalt, zur De-Anonymisierung einer Person sehr gut genutzt werden kann.
- Die zentrale Implikation ist die Folgende: Um was für eine zweite Datenquelle es sich handelt, das ist nicht wichtig. Es könnte genauso gut eine Datenbank sein, die durch das Crawlen[‡] von Facebook oder durch einen einfachen Datenleak erstellt wurde. Entscheidend ist, dass eine kleine Datenmenge ausreicht, um die große Datenbank zu de-anonymisieren.

Aus meiner Sicht haben die Ergebnisse von Narayanan und Shmantikov folgende Implikationen für medizinische Daten:

- Intuitiv kann man sich vorstellen, dass jede Bewertung eines Filmes durch einen Kunden ein Punkt einer sehr großen Datenwolke darstellt. Überraschenderweise identifizieren sehr wenige Punkte einen Kunden. Da der Informationsgehalt medizinischer Daten ungleich höher ist als der von Filmbewertungen, ist davon auszugehen, dass die Anonymisierung dieser Daten auch deutlich schwieriger ist. Überträgt man die Analyse von Narayanan und Shmantikov zum Beispiel auf die Medikamente eines Patienten, so reicht wahrscheinlich die Information über ein paar wenige Medikamente eines Patienten aus, um diesen eindeutig zu identifizieren.
- Die Techniken von Narayanan und Shmantikov sind elegant und überraschend einfach. Um die De-anonymisierungsangriffe durchzuführen, benötigen die Autoren weder eine Vielzahl von externen Quellen noch Hochleistungscomputer. Folglich können die Angriffe ohne großen Aufwand auf andere Anwendungsgebiete übertragen werden. Beispielsweise könnte die Information einer bestimmten Operation an einem bestimmten Tag mit einer geteilten Krankengeschichte auf Instagram, Twitter, PatientsLikeMe oder einem Datenleck in einer Health-App verknüpft werden.

[‡]Unter einem Crawler versteht man ein Programm, welches die Inhalte einer Website selbstständig durchsucht und dieInformation vollständig ausliest.

2.4.2 Der Fall der Datenspende App des RKIs

Das zweite Fallbeispiel betrifft die Corona-Datenspende-App des Robert-Koch-Institutes (RKI) [13]. Bei dieser App können die Bürgerinnen und Bürger freiwillig Daten spenden mit dem Ziel, eine Fieberkurve zu berechnen. Mittels dieser Fieberkurve soll dann „die Zahl an COVID-19 Erkrankten abgeschätzt und „die Entstehung neuer COVID-19-"Hot Spots" sichtbar" [14] gemacht werden. Laut der Seite des RKIs, spendeten über 500.000 Menschen über 400.000.000 Daten[§]. Um die App zu nutzen, stellen die Benutzer:innen folgende grundlegende Informationen zur Verfügung:

- Postleitzahl
- Größe (in Schritten à 5 cm)
- Gewicht (in Schritten à 5 kg)
- Alter (in Schritten à 5 Jahre)

Neben diesen allgemeinen Daten spenden die Nutzer:innen verschiedene Gesundheitsdaten: Aktivitäts- und Ruhepuls, Schritte, Kalorienverbrauch, zurückgelegte Entfernung, die Anzahl an gestiegenen Treppen und Informationen über den Schlaf [5]. Mittels dieser Daten berechnet das RKI dann eine Fieberkurve, wie sie in Abbildung 2.1 dargestellt ist. Das RKI versprach den Nutzern der Datenspende App, dass die Daten pseudonymisiert

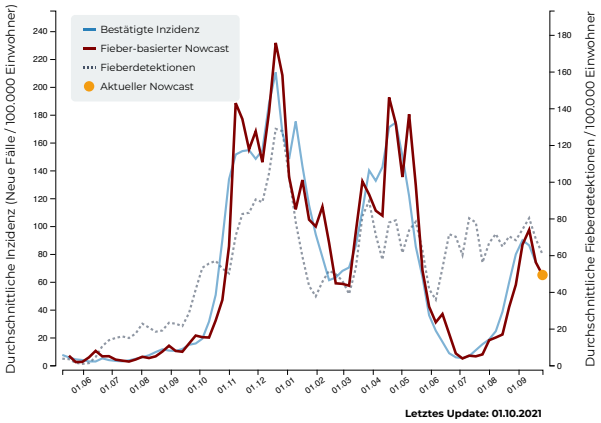


Abb. 2.1: Fieberkurve des RKIs [15].

gespeichert werden und keine Rückschlüsse auf das Individuum möglich sind.

Re-Identifikationsangriffe

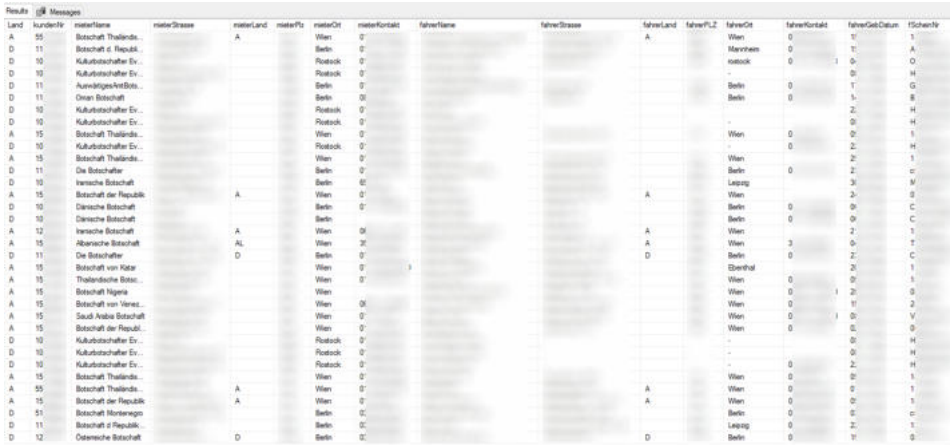
Das RKI wendet verschiedene Techniken zur Pseudonymisierung der Daten an, die bereits im Abschnitt 2.3 vorgestellt wurden: Es wird kein Identifikator, wie zum Beispiel

[§]Die Anzahl der tatsächlichen Spender ist laut RKI deutlich höher. Die Zahl auf der Website entspricht der Anzahl an Spender:innen, die ein vollständiges Profil hinterlegt haben.

die Nummer des Personalausweises, gespeichert (Informationsunterdrückung). Darüberhinaus werden die Daten, wie die Größe, das Gewicht und das Alter in Abschnitten von 5 Einheiten gespeichert (Generalisierung).

In einer gemeinsamen Forschungsarbeit mit Pascal Berrang konnten wir jedoch zeigen, dass diese Techniken nicht ausreichen und eine eindeutige Zuordnung in sehr vielen Fällen ohne großen Aufwand möglich ist [16]. Unser Angriff ist nicht spezifisch für die Datenspende App und kann auf beliebige Anwendungen übertragen werden. In dem Angriff gehen wir davon aus, dass der Angreifer Zugriff auf zwei Datenbanken hat. Die erste Datenbank speichert soziodemographische Daten und auch den Namen eines Benutzers. In unserer Analyse wird von der Existenz dieser Datenbank ausgegangen. Diese kann der Angreifer zum

Beispiel durch eine der zahlreichen Datenlecks oder gestohlenen Datensätze aufgebaut haben. Eines dieser Datensätze könnte zum Beispiel die vollständige Kundendatei von Buchbinder sein, einem der größten deutschen Autovermieter im Privatkundensegment [17]. Diese Daten enthielten Fahrer mit Namen, Adresse, Geburtsdatum, Führerscheinnummer und -Ausstellungsdatum und konnten freizugänglich über Internet abgerufen werden. Ein Ausschnitt der Datenbank ist in Abbildung 2.2 dargestellt.



Land	kundenNr	mieterName	mieterStrasse	mieterLand	mieterPlz	mieterOrt	mieterKontakt	fahrerName	fahrerStrasse	fahrerLand	fahrerPLZ	fahrerOrt	fahrerKontakt	fahrerGebDatum	fSchein Nr
A	55	Botschaft Thailandis...		A		Wien	0			A		Wien	0	1	1
D	11	Botschaft d. Republi...				Berlin	0					Mannheim	0	1	A
D	10	Kulturbotschafter Ev...				Rostock	0					rostock	0	0	O
D	10	Kulturbotschafter Ev...				Rostock	0					-		0	H
D	11	AuswärtigesAmtBots...				Berlin	0					Berlin	0	1	G
D	11	Oman Botschaft				Berlin	0					Berlin	0	1	B
D	10	Kulturbotschafter Ev...				Rostock	0							2	H
D	10	Kulturbotschafter Ev...				Rostock	0					-		0	H
A	15	Botschaft Thailandis...				Wien	0					Wien	0	0	1
D	10	Kulturbotschafter Ev...				Rostock	0					-	0	2	H
A	15	Botschaft Thailandis...				Wien	0					Wien		2	1
D	11	Die Botschafter				Berlin	0					Berlin	0	2	c
D	10	Iranische Botschaft				Berlin	0					Leipzig		3	M
A	15	Botschaft der Republik		A		Wien	0			A		Wien		2	0
D	10	Dänische Botschaft				Berlin	0					Berlin	0	0	C
D	10	Dänische Botschaft				Berlin						Berlin	0	0	C
A	12	Iranische Botschaft		A		Wien	0			A		Wien		2	1
A	15	Albanische Botschaft		AL		Wien	3			A		Wien	3	0	T
D	11	Die Botschafter		D		Berlin	0			D		Berlin	0	2	C
A	15	Botschaft von Katar				Wien	0					Eberthal		2	1
A	15	Thailandische Botsc...				Wien	0					Wien	0	0	1
A	15	Botschaft Nigeria				Wien						Wien	0	2	0
A	15	Botschaft von Venez...				Wien	0					Wien	0	1	J
A	15	Saudi Arabia Botschaft				Wien	0					Wien	0	0	V
A	15	Botschaft der Republ...				Wien	0					Wien	0	0	0
D	10	Kulturbotschafter Ev...				Rostock	0					-		0	H
D	10	Kulturbotschafter Ev...				Rostock	0					-		0	H
D	10	Kulturbotschafter Ev...				Rostock	0					-	0	2	H
A	15	Botschaft Thailandis...				Wien	0					Wien	0	0	1
A	55	Botschaft Thailandis...		A		Wien	0			A		Wien	0	0	1
A	15	Botschaft der Republik		A		Wien	0			A		Wien	0	0	1
D	51	Botschaft Montenegro				Berlin	0					Berlin	0	0	c
D	11	Botschaft d Republik...				Berlin	0					Leipzig	0	2	1
D	12	Österreiche Botschaft		D		Berlin	0			D		Berlin	0	1	0

Abb. 2.2: Ausschnitt der ungeschützten Datenbank von Buchbinder [17].

Die zweite Datenbank enthält anonymisierte Daten. Der Angreifer versucht nun die Einträge zu verknüpfen und damit eine Re-Identifikation der Benutzer in der zweiten Datenbank zu erreichen. Findet der Angreifer eine eindeutige Abbildung zwischen den Datenbanken, dann gibt er diese aus. Kommen mehrere Einträge in Frage, dann wählt er einen der möglichen Einträge zufällig aus. Des Weiteren vergleichen wir die Erfolgswahrscheinlichkeit des Angreifers basierend auf drei unterschiedlichen Hintergrundinformationen, also zusätzliches Wissen, dass der Angreifer sich vor seinem Angriff aneignete. Im ersten Fall (a) nehmen wir an, dass der Angreifer keine weitere Hintergrundinformation hat. Im zweiten Fall (b) weiß der Angreifer, dass die Nutzer eine Smartwatch tragen und im dritten Fall (c) weiß der Angreifer, dass die Teilnehmer die Datenspende App nutzen. Die Ergebnisse der durchschnittlichen Anyomitätsmengen sind in Abbildung 2.3

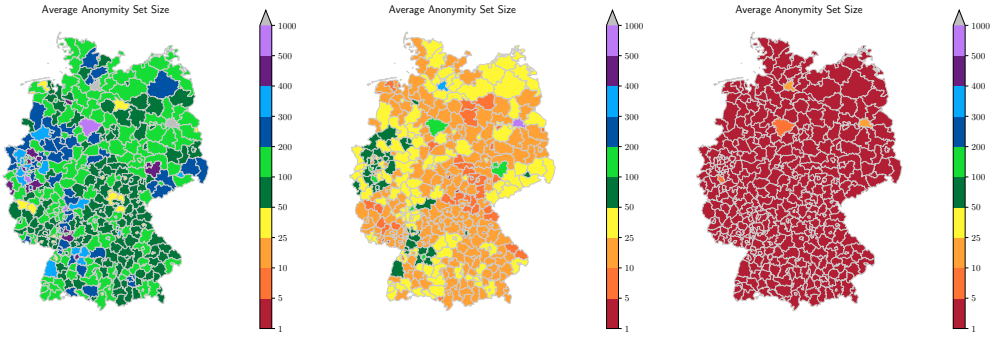


Abb. 2.3: links: (a) keine Hintergrundinformation, Mitte (b) die Nutzer verwenden eine Smartwatch, rechts die Nutzer verwenden die Datenspende App [16].

dargestellt. Die Farben drücken die Größe der Anyomitätsmengen aus, wobei die Farbe lila einer hohen, grün und gelb einer mittleren und rot gar keiner Anyomitätsmenge entspricht. An den Darstellungen wird deutlich, dass große Anonymitätsmenge bestenfalls in den großen Städten wie Berlin und Hamburg erreicht werden. Außerdem sieht man im mittleren Bild, dass eine scheinbar harmlose Information wie „Smartwatch Nutzer“ die Anomyitätsmenge drastisch reduziert. Kumuliert man alle Ergebnisse in einer sogenannten Verteilungsfunktion, dann erhält man die Kurven, welche in Abbildung 2.4 dargestellt sind. Die blaue stellt die Anomyitätsmenge ohne Anwendung jeglicher Hintergrundin-

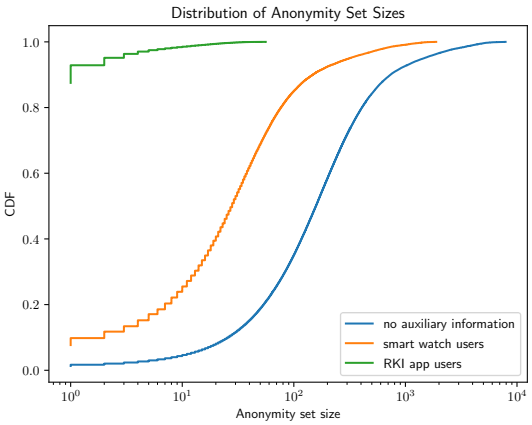


Abb. 2.4: Verteilungsfunktionen: blau entspricht keiner Hintergrundinformation, orange bedeutet dass der Angreifer weiß, dass die Nutzer eine Smartwatch verwenden und bei grün wird das Wissen, dass der Nutzer die Datenspende App verwendet mit einbezogen. [16].

formation dar. Man sieht, dass **über 10% der Population** eine Anomyitätsmenge von weniger als 40 hat und dass lediglich 10% der Population eine Anomyitätsmenge von mindestens 300 hat.

Hat der Angreifer das Zusatzwissen „Smartwatch Nutzer“, so sehen wir im mittleren Bild, dass **30% der Population eine Anomyitätsmenge von höchstens 30** hat und lediglich 15% kommen auf eine Anomyitätsmenge von über 100.

Noch dramatischer wird es falls der Angreifer weiß, dass ein Nutzer die Datenspende App verwendet. In diesem Fall sind **über 87% der Population eindeutig identifizierbar**.

Interpretation der Ergebnisse

Die Ergebnisse unserer Forschungsarbeit sind ein weiteres Beispiel dafür, dass sehr wenige Informationen ausreichen, um eine Person eindeutig zu identifizieren. Der Angriff ist erneut sehr einfach und kann mit einfachen Mitteln innerhalb kürzester Zeit umgesetzt werden. Es ist davon auszugehen, dass der Prozentsatz noch verbessert werden kann, wenn man weitere Informationen und Datenquellen hinzuzieht.

2.5 Zusammenfassung der Ergebnisse

Das Ziel dieses Kapitels ist die Beantwortung der Frage, wie leicht pseudonymisierte (Gesundheits-)Daten re-personalisiert werden können. Zu diesem Zweck wurden zwei verschiedene Forschungsergebnisse vorgestellt, das erste beschäftigte sich mit der Re-Personalisierung von pseudonymisierten Nutzerdaten von Netflix und das zweite mit der Re-Personalisierung von pseudonymisierten Nutzerdaten der Corona-Datenspende-App des RKIs. Die vorgestellten Angriffe sind technisch sehr einfach und zeigen, dass jede Information wie zum Beispiel „Smtartwatch Nutzer“ große Auswirkungen auf den Anonymisierungsgrad haben. Diese vorgestellten Angriffe zeigen deutlich, dass einfache Techniken zur Anonymisierung nicht ausreichen. An dieser Stelle möchte ich betonen, dass diese zwei Forschungsarbeiten keine Ausnahme darstellen und dass es sich nicht um überraschende neue Erkenntnisse handelt. Vielmehr zeigt insbesondere das zweite Beispiel, dass aus den Fehlern der Vergangenheit nicht gelernt wurde. In der Praxis herrscht immer noch der Irrglaube, dass die Entfernung des Namens die Identität des Einzelnen ausreichend schützt.

In Bezug auf die ursprüngliche Fragestellung, wie leicht pseudonymisierte (Gesundheits-) Daten re-personalisiert werden können, komme ich zu dem Schluss, dass dies ohne großen Aufwand möglich ist[¶]. Das wesentliche Problem besteht darin, dass die gesammelten Daten sehr feingranular sind und diese Feingranularität führt dazu, dass einzelne Personen leicht Re-Identifiziert werden können. In § 3 Abs. 1 Nr. 3 lit. b aa DaTraV wird zum Beispiel festgelegt, dass die genaue Medikation der Arzneimittel aufgeführt wird, inklusive dem Mengenfaktor, Datum der Abgabe durch die Apotheke und Institutionskennzeichen der abgebenden Apotheke. Diese Daten spannen einen deutlich größeren Raum auf als das einfache Beispiel von den Filmbewertungen im Falle von Netflix. Folglich wird die Re-Identifikation eines Individuums auch deutlich einfacher werden. Um diese Einschätzung zu untermauern, rufe ich das Beispiel der öffentlichen Datenbank von Buchbinder [17]

[¶] In Abschnitt 3.2 werden fortgeschrittene Techniken zum Schutz der Privatsphäre beschrieben, die einen Schutz bieten.

in Erinnerung und verknüpfe diese Daten mit den pseudonymisierten Gesundheitsdaten. Wie viele Menschen wird es zum Beispiel in München geben, die 30 Jahre alt sind, in einem bestimmten Postleitzahlbereich leben und aufgrund einer Gehbehinderung auf einen Rollstuhl angewiesen sind? Autovermietungen bieten speziell für diese Menschen ebenfalls Autos an. Von den pseudonymisierten Gesundheitsdaten wissen wir, dass es in München in dem besagten PLZ Bereich eine Person gibt, die spezielle Medikamente für Menschen mit einer bestimmten Gehbehinderung erhält. Aufgrund der Daten wissen wir auch, wann die Person dieses Medikamente immer in der gleichen Apotheke abgeholt, die sich in der Nähe des Wohnortes gemäß der Datenbank von Buchbinder befindet. Verknüpft man nun diese Daten, so kommt man schnell zu dem Schluss, dass es sich um die Daten von Peter Müller handeln muss.

Auch wenn dies ein hypothetisches Beispiel ist, so zeigt dies deutlich, dass die vorhandenen Informationen die Anonymitätsmenge stark einschränken und höchstwahrscheinlich zur eindeutigen Identifikation der Patienten führen. Folglich ist davon auszugehen, dass das „schlüsselabhängige Verfahren zur Pseudonymisierung“, welches keine Rückschlüsse „auf das Lieferpseudonym oder die Identität des Versicherten“ erlaubt, wie es § 303 c Abs. 2 SGB V vorschreibt, keinen Schutz bietet, da die Re-Identifikation über die anderen Merkmale in den Daten erfolgt.

3 Alternative Ansätze zur Bereitstellung von (lediglich) pseudonymisierten Datensätzen

Fragestellung: Welche Alternativen gibt es zur Bereitstellung von (lediglich) pseudonymisierten Datensätzen und sind sie von vergleichbarer Qualität (gestaffelt nach Forschungszweck)?

In diesem Kapitel werden alternative Ansätze zur Bereitstellung von pseudonymisierten Datensätzen vorgestellt. Falls die Verfahren keine oder nur wenig mathematische Grundlagen erfordern, werden diese an einfachen Beispielen erläutert. Des Weiteren werden die Vor- und Nachteile der Verfahren erläutert. Die bekannten Ansätze lassen sich aus meiner Sicht in drei große Gruppen unterteilten. Die erste Gruppe umfasst die Verfahren, welche die Daten nicht verfälschen und lediglich sicher stellen, dass „genug“ Möglichkeiten infrage kommen. Die zweite Gruppe verrauscht die Daten, d. h., es werden gezielt kleine Fehler eingefügt und die dritte Gruppe führt Berechnungen auf verschlüsselten Daten durch.

3.1 K-Anonymität und verwandte Ansätze

In diesem Abschnitt wird das Verfahren k-Anonymität und deren verwandte Ansätze vorgestellt. Die wesentliche Idee dieser Gruppe an Verfahren besteht darin, dass der Zugriff nur auf Daten ermöglicht wird, bei denen stets mehrere Personen in Frage kommen, da diese die gleichen Eigenschaften haben.

3.1.1 K-Anonymität

Der erste Ansatz ist bekannt unter dem Namen „k-Anonymität“ (*engl. k-anonymity*) und folgt der Idee, dass mindestens k Datenbankeinträge bei dem Versuch der Re-Personalisierung infrage kommen [18]. Zum besseren Verständnis wird das Verfahren anhand des Beispiels aus Unterabschnitt 2.3.1 verdeutlicht. Nachdem die Identifikatoren entfernt und das Prinzip der Generalisierung angewendet wurde, erhielten wir Tabelle 3.1.

Diese Tabelle wird in sogenannte Äquivalenzklassen unterteilt. Unter einer Äquivalenzklasse versteht man eine Menge von Elementen, die sich „gleichen“, d.h., bei denen bestimmte Eigenschaften gleich sind. In Tabelle 3.1 sind drei Äquivalenzklassen dargestellt. Zur einfacheren Betrachtung wurde jeder Äquivalenzklasse eine Farbe zugewiesen. Die

Alter	Geschlecht	PLZ	Diagnose
[50-59]	M	38***	Krebs
[50-59]	M	38***	Grippe
[40-49]	W	91***	Krebs
[40-49]	W	91***	Krebs
[40-49]	W	91***	COVID-19
[70-79]	M	23***	COVID-19
[70-79]	M	23***	COVID-19

Tab. 3.1: Äquivalenzklassen der Tabellen

Äquivalenzklassen umschließen jeweils Personen, die im gleichen Altersbereich liegen, das gleiche Geschlecht haben und im gleichen Postleitzahlbereich leben. Wie man farblich sehen kann, enthält jede Äquivalenzklasse mindestens zwei Elemente, folglich erreicht die anonymisierte Tabelle 2-Anonymität. Der Wert $k = 2$ ist der kleinste Wert, um eine unmittelbare Identifizierung auszuschließen. Intuitiv sollte ein höherer Wert von k das Risiko einer Identifikation reduzieren (siehe Unterabschnitt 3.1.4 bezüglich der generellen Schwächen des Ansatzes). Im biomedizinischen Bereich wird ein Wertebereich von $k = 3$ bis $k = 25$ als normal angesehen und in der Regel ein Wert von $k = 5$ verwendet [19].

3.1.2 Schwächen von K-Anonymität

In diesem Abschnitt werden die Schwächen dieses Ansatzes beschrieben.

Homogenitätsangriffe

Der Begriff der k-Anonymität stellt lediglich sicher, dass es *mindestens* k Elemente in den jeweiligen Äquivalenzklassen gibt. Er stellt jedoch nicht sicher, dass es eine gewisse Varianz in den sensitiven Attributen gibt. Dies führt dazu, dass man sensitive Attribute durchaus bestimmten Personen zuordnen kann. Solche Angriffe werden als Homogenitätsangriffe bezeichnet und ein entsprechendes Beispiel ist in Tabelle 3.1 gegeben. In der blauen Äquivalenzklasse wird deutlich, dass alle Teilnehmer*Innen in der Klasse an COVID-19 erkrankt sind. Dies ist bei den anderen beiden Klassen nicht der Fall, da es dort mehrere Möglichkeiten für eine bestimmte Diagnose gibt. Das bedeutetet, dass die Kenntnis davon, dass eine bestimmte Person zwischen 70 und 79 Jahre alt ist, genügt, um auf ihre Erkrankung zu schließen. Diese Angriffe können beliebig erweitert werden: Weiß zum Beispiel eine Person A, dass die Person B im PLZ Bereich 23*** lebt und 74 Jahre alt ist, so kann erneut auf die Krankheit geschlossen werden.

3.1.3 l-Diversität

Im Unterabschnitt 3.1.2 wurde als Schwäche von k-Anonymität so genannte „Homogenitätsangriffe“ identifiziert. Bei dieser Klasse von Angriffen haben alle Einträge einer

Äquivalenzklasse dasselbe sensible Attribut. Anhand des Beispiels aus Tabelle 3.1 wird deutlich, dass alle Patienten im Alter von 70-80 an COVID-19 erkrankt sind. Gelingt es nun dem Angreifer eine Person zu dieser Äquivalenzklasse zuzuordnen, so kennt dieser automatisch die Diagnose des Patienten.

Ein weiteres Problem entsteht durch Zusatzwissen (sog. *Background Knowledge* Angriffe), welches der Angreifer über den Datensatz hat (siehe Unterabschnitt 2.4.2). Mittels dieses Hintergrundwissens kann der Angreifer einzelne Elemente ausschließen und somit auf die eigentliche Krankheit schließen. Um das Problem zu verdeutlichen betrachten wir erneut das Beispiel in Tabelle 3.1. Wir nehmen an, dass der Angreifer weiß, dass es sich um eine weibliche Person handelt. Folglich ist die Person entweder an Krebs oder an COVID-19 erkrankt. Handelt es sich zum Beispiel um eine Kollegin, die einen negativen COVID-19 Test vorgelegt hat, so ist klar, dass die Kollegin höchstwahrscheinlich an Krebs erkrankt ist.

Um diese Art von Angriffen zu umgehen, wurde von Machanavajjhala et al. das Konzept der *l-Diversität* (engl. *l-diversity*) entwickelt [20]. Der Parameter l gibt ein Maß für die Varianz innerhalb der Äquivalenzklasse an. Das einfachste Beispiel beschreibt die „*verschiedene* 1-Diversität“, diese besagt, dass es mindestens einen unterschiedlichen Wert für sensible Attribute geben muss. Natürlich ist das Maß „*verschiedene* 1-Diversität“ relativ schwach und kann leicht mit statistischen Angriffen umgangen werden, indem die Verteilung der Daten innerhalb einer Äquivalenzklasse mit der repräsentativen Verteilung der Gesamtbevölkerung abgleicht.

Um solche Probleme ebenfalls zu umgehen, wurden verschiedene Maße vorgeschlagen, wie zum Beispiel ein Entropie-basierter Ansatz oder ein rekursiver Ansatz der l-Diversität [20].

Ein Beispiel für eine Datenbank mit einem Diversitätswert von $l = 3$ ist in Tabelle 3.2 dargestellt. Der Wert $l = 3$ wird erreicht da in jeder Äquivalenzklasse mindestens 3 unterschiedliche Diagnosen vorhanden sind.

Alter	Geschlecht	PLZ	Diagnose
[50-59]	M	38***	Krebs
[50-59]	M	38***	Grippe
[50-59]	M	38***	Diabetes
[50-59]	M	38***	COVID-19
[40-49]	W	91***	Krebs
[40-49]	W	91***	Diabetes
[40-49]	W	91***	COVID-19
[70-79]	M	23***	Diabetes
[70-79]	M	23***	Krebs
[70-79]	M	23***	COVID-19

Tab. 3.2: Datenbank mit einem Diversitätswert von $l = 3$.

3.1.4 Weitere Ansätze

Im vorherigen Abschnitt wurde bereits die Schwäche des Ansatzes der l-Diversität in Bezug auf statistische Analyse diskutiert. Um die Effektivität dieser Angriffe zu reduzieren, wurde der Begriff der *t-closeness* entwickelt [21]. Intuitiv wird bei diesem Ansatz versucht, nur Daten zu veröffentlichen, deren statistische Verteilung ungefähr mit der Verteilung der Bevölkerung (bzw. passend für die konkrete Anwendung) übereinstimmt oder dieser sehr nahekommt. Technisch kann die Abweichung zweier Wahrscheinlichkeitsverteilungen zum Beispiel durch die sogenannte „Wasserstein Metrik“ bestimmt werden. Diese ist jedoch mathematisch sehr anspruchsvoll und wird an dieser Stelle nicht weiter erläutert. Beispiele findet man in dem Artikel „Für immer anonym: Wie kann De-Anonymisierung verhindert werden?“ von Jäschke et al.[19].

3.1.5 Allgemeine Probleme bei der Pseudonymisierung und Anonymisierung von Datenbanken

Unabhängig von den eingesetzten Verfahren ergeben sich verschiedene Probleme, die im Folgenden kurz skizziert werden.

Umsortierung der Einträge

Die Tabellen 2.3 und 3.1 sind inhaltlich identisch, jedoch wurden die Einträge in Tabelle 3.1 gemäß der Äquivalenzklassen umsortiert. Generell ist die (zufällige) Permutation der Einträge wichtig, da die Position der Einträge ebenfalls Informationen über den Eintrag liefert und eine Re-Personalisierung sonst einfacher möglich ist.

Unterschiedlich anonymisierte Tabellen

Ein weiteres Problem tritt auf, wenn gleiche Tabellen zu unterschiedlichen Zwecken unterschiedlich anonymisiert werden. Dieses Problem kann in der Praxis leicht auftreten, da oftmals redundante Datensätze von unterschiedlichen Stellen gespeichert werden. Beispielsweise speichern Krankenkassen, Ärzte und Krankenhäuser viele Daten eines Patienten doppelt. Werden diese Daten nun auf unterschiedliche Arten anonymisiert, können durch einfache Kombination der Tabellen Teile der Datenbank wieder hergestellt werden. Durch diese Kombination ist oftmals die Zuordnung eines Eintrages zu einer Person möglich. Beispiele für solche Rekonstruktionsangriffe sind in Unterabschnitt 2.4.1 und Unterabschnitt 2.4.2 gegeben.

Dynamische Tabellen

In der Regel sind medizinische Datensätze nicht statisch, sondern verändern sich über Zeit, wenn zum Beispiel neue Befunde hinzugefügt werden. Durch die Modifikationen verändert sich die Tabelle ständig und es treten ähnliche Probleme auf wie im vorherigen Abschnitt beschrieben. Bei allen Modifikationen muss zu jedem Zeitpunkt sichergestellt sein, dass die k-Anonymität erhalten bleibt.

3.1.6 Vor- und Nachteile dieser Ansätze

Das Charmante an diesen Ansätzen ist, dass sie leicht verständlich sind. Da für die Implementierung keine tieferen mathematischen Grundkenntnisse notwendig sind, können diese Ansätze mit relativ geringem Aufwand umgesetzt werden.

Die wesentliche Schwäche dieser Ansätze besteht aus meiner Sicht in den trügerischen Garantien. Welches Hintergrundwissen der Angreifer hat, ist zum Zeitpunkt der Anonymisierung schwer bzw. unmöglich abschätzbar. Werden im Zuge der Anonymisierung Daten entfernt, so ist die Qualität der Evaluierung je nach Anwendung fraglich, da entfernte sensitive Attribute die Aussage einer Evaluation entscheidend verändern können.

3.2 Differential Privacy

In diesem Abschnitt wird ein alternativer Ansatz, der unter dem Namen „Differential Privacy“ (auf deutsch „Differenzielle Privatsphäre“) bekannt ist, vorgestellt [22]. Die wesentliche Idee dieses Ansatzes besteht darin, dass die Daten gezielt verrauscht werden, d. h., es werden Fehler eingefügt, sodass das Ergebnis einer Berechnung nur minimal verfälscht wird, jedoch die Privatsphäre eines Einzelnen geschützt wird.

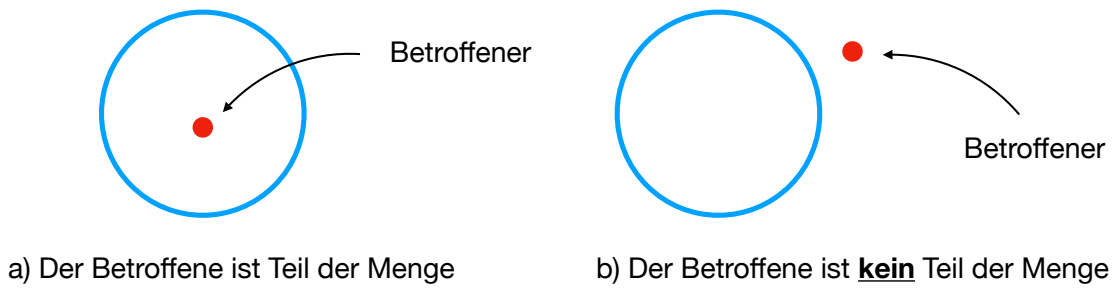


Abb. 3.1: Visualisierung der wesentlichen Idee hinter Differential Privacy. Das Ergebnis der Auswertung auf den Daten soll in beiden Fällen zu vergleichbaren Ergebnissen führen.

Diese grundlegende Idee ist in Abbildung 3.1 visualisiert. Im Teil (a) ist eine Menge an Personen zu sehen und ein Betroffener (roter Kreis) ist Teil dieser Menge. Im rechten Teil ist die gleiche Menge abgebildet, jedoch befindet sich der Betroffene nicht in der Menge. Eine Berechnung auf beiden Mengen führt jedoch zu vergleichbaren Ergebnissen, weil der eingeführte Fehler zwar die Anonymität des Einzelnen schützt, sich jedoch nur minimal auf das Ergebnis der Berechnung auswirkt. Dies bedeutet insbesondere, dass die Privatsphäre des Betroffenen geschützt ist, da die Daten des Betroffenen nicht zu der Berechnung des Ergebnisses beitragen. Würden die Daten die Berechnung signifikant verändern, dann gäbe es einen messbaren Unterschied zwischen der Berechnung auf beiden Mengen, was der Grundeigenschaft von Differential Privacy widerspricht.

Auf den ersten Blick erscheint das Versprechen von Differential Privacy als zu gut um wahr zu sein. Wenn die Daten des Betroffenen nicht das Ergebnis verändern können, dann dürfte doch niemand etwas zu den Daten beitragen, andernfalls sollte es eine signifikante

Veränderung in dem Ergebnis ergeben. Um diesen scheinbaren Widerspruch aufzulösen, muss man sich über zwei Grundsätze dieses Ansatzes im Klaren sein:

- Differential Privacy erreicht die Garantien, indem Teile der Daten verrauscht werden. Durch diese Operation kann es passieren, dass die Daten eines Betroffenen *verfälscht* zu dem Endergebnis beitragen.
- Die Technik funktioniert besonders gut bei großen Datenmengen, da die Fehler, die durch die Verrauschung der Daten eingefügt werden, sich nur geringfügig auf das Gesamtergebnis auswirken. Gleichzeitig wurde jeder Eintrag jedoch mit einer gewissen Wahrscheinlichkeit verfälscht, sodass man sich bei einem einzelnen Datum nicht sicher sein kann, dass dies der Realität entspricht.

Generell gilt es bei diesem Ansatz eine Balance zwischen dem Schutz der Privatsphäre eines Betroffenen und dem Nutzen des Ergebnisses (der sogenannten *utility*) zu finden. Das vollständige Verrauschen der Daten führt zu dem höchsten Schutz der Privatsphäre, jedoch sind die Ergebnisse damit wertlos. Auf der anderen Seite liefern unverfälschte Daten die besten Ergebnisse, jedoch schützen diese die Privatsphäre nicht. Abhängig von der Art der Berechnung auf den Daten sind verschiedene Techniken zum Verrauschen der Daten bekannt, die ein gewisses Optimum zwischen dem Schutz der Privatsphäre und dem Nutzen des Ergebnisses liefern.

Im Folgenden werden zwei Ansätze aus dem Bereich der Differential Privacy vorgestellt, die sich in dem Vertrauensmodell stark unterscheiden. Im Anschluss wird eines dieser Modelle an einem praktischen Beispiel der Datenspende App des RKI's (im Unterabschnitt 2.4.2 werden die Re-Personalisierungsangriffe auf das System vorgestellt), beschrieben. Die Vor- und Nachteile dieses generellen Ansatzes werden im Unterabschnitt 3.2.3 diskutiert.

3.2.1 Differential Privacy Zentralisierter Ansatz

Das grundlegende Setting des *zentralisierten* Ansatzes ist in Abbildung 3.2 dargestellt. In diesem Ansatz senden die Nutzer des Systems die Daten *unverfälscht* an eine zen-

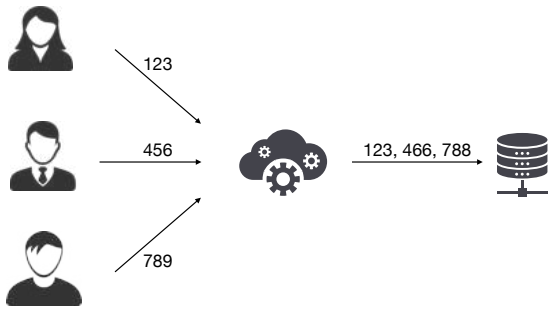


Abb. 3.2: Darstellung des zentralisierten Ansatzes von Differential Privacy.

trale Stelle. In dem dargestellten Beispiel senden die Teilnehmer die Werte 123, 456

und 789. Das explizite Senden der Daten ist nicht der entscheidende Punkt, sondern lediglich die Tatsache, dass diese Stelle direkten Zugriff auf alle Daten erhält. Folglich kann es sich auch um eine Datenbank handeln, die dort gespeichert wird. Die zentrale Stelle wird somit als vertrauenswürdig angesehen. Vertrauenswürdig bedeutet in diesem Zusammenhang, dass die zentrale Partei die Daten nicht herausgibt und gemäß den festgelegten Richtlinien verrauscht. Der Angreifer erhält somit keinen Zugriff auf diese Stelle, sondern darf Berechnungen auf den verrauschten Daten ausführen. Zum Schutz der Privatsphäre der Teilnehmer werden die Daten nun verrauscht, d. h., bei jeder Stelle der Zahl wird mit einer bestimmten Wahrscheinlichkeit ein kleiner Fehler eingefügt. In dem einfachen Beispiel aus Abbildung 3.2 wird der erste Wert unverfälscht gespeichert, die beiden anderen Werte werden jedoch leicht verändert. Nach dem Verrauschen der Daten können nun Berechnungen auf der Datenbank durchgeführt werden. Ist man nun an dem Durchschnittswert interessiert, dann würde dieser in der unverfälschten Variante $(123+456+789)/3 = 456$ betragen. Führt man die gleiche Berechnung nun auf den verrauschten Werten aus, so ergibt dies den Wert $(123 + 466 + 788)/3 = 459$, der minimal von dem eigentlichen Wert abweicht.

Erlangt ein Angreifer nun Zugriff auf die Datenbank, so kann sich dieser nicht sicher sein, ob die erste Teilnehmerin wirklich den Wert 123 beigetragen hat oder ob dieser Wert verrauscht wurde. Die Wahl des einzufügenden Fehlers hängt von den Daten und der Anwendung ab. Generell lässt sich jedoch sagen, dass typische Anwendungsgebiete, wie zum Beispiel herkömmliche statistische Auswertungen, in der Forschung sehr gut verstanden sind und entsprechende Optima zwischen dem Schutz der Privatsphäre eines Individuums und dem Nutzen des Ergebnisses [22] gefunden wurden oder aus bekannten Ergebnissen leicht ableitbar sind.

3.2.2 Differential Privacy Dezentraler Ansatz

Im Gegensatz zu dem zentralisierten Ansatz gibt es verschiedene *dezentrale Ansätze*, bei denen es keine zentrale vertrauenswürdige Entität gibt, wie zum Beispiel das so genannte RAPPOR Framework [23]. Dies impliziert auch, dass die Daten direkt an der Quelle verrauscht werden müssen bevor diese in der Datenbank gespeichert werden. Der Angreifer erhält im Anschluss Zugriff auf die Datenbank.

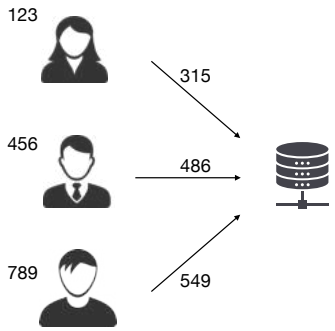


Abb. 3.3: Darstellung des dezentralen Ansatzes von Differential Privacy.

Ein vereinfachtes Beispiel ist in Abbildung 3.3 dargestellt. Die Daten werden direkt bei den Parteien erfasst, genau wie in Abbildung 3.2 sind die unverfälschten Werte der drei Personen 123, 456 und 789. Jeder Teilnehmer ändert jede Stelle mit einer bestimmten Wahrscheinlichkeit und sendet die verrauschten Werte dann an den Server bzw. an die Datenbank. In diesem Beispiel werden die Werte 315, 486 und 549 gesendet. Wird nun wie im vorherigen Beispiel der Durchschnittswert berechnet, so ergibt dies in diesem Fall den Wert $(315 + 486 + 549)/3 = 450$.

3.2.3 Herausforderungen, Vor- und Nachteile dieser Ansätze

Im Gegensatz zu den Techniken aus Abschnitt 3.1 können durch Verwendung von Techniken aus dem Bereich der Differential Privacy einfache Linking/Verknüpfungsangriffe ausgeschlossen werden. Gleichzeitig ist die Anwendung dieser Technik jedoch ungleich komplexer. Beispielsweise macht es einen Unterschied, ob die Datenbank einmal aufgesetzt und verrauscht wurde, oder ob kontinuierlich neue Daten eingefügt werden. Des Weiteren muss festgelegt sein wie genau die Daten verrauscht werden und wie der Angreifer auf die Daten zugreifen kann. Werden zum Beispiel die Daten immer wieder aufs Neue verrauscht, dann kann der Angreifer den ursprünglichen Wert mit hoher Wahrscheinlichkeit ermitteln indem er immer wieder die gleiche Frage stellt und dann ein Mehrheitsvotum über die Antworten bildet.

Vor- und Nachteile

Auf den ersten Blick erscheint der dezentrale Ansatz als die vorzuziehende Lösung. Dies stimmt jedoch nur bedingt, da die Umsetzung ungleich schwieriger ist. Bei dem zentralen Ansatz hat der Algorithmus einen Überblick über alle Daten und kann diese im Gesamten verrauschen. Dies ist bei der dezentralen Lösung nicht möglich, da jeder Teilnehmer unabhängig voneinander die Daten verrauscht. Folglich erhält man bei einem zentralen Ansatz oftmals bessere Ergebnisse, die sich auch leichter umsetzen lassen.

3.2.4 Die Datenspende App des RKIs: Praktisches Beispiel von Differential Privacy

In diesem Abschnitt wird die Anwendbarkeit des dezentralen Ansatzes von Differential Privacy an einem praktischen Beispiel vorgestellt. In diesem Beispiel wird erneut die Datenspende App des RKIs aufgegriffen und gezeigt, dass die gleiche Funktionalität, also die Berechnung der Fieberkurve, realisiert werden und gleichzeitig die Linking Angriffe (siehe Unterabschnitt 2.4.2) nicht mehr möglich sind [16]. Da es sich bei der Berechnung der Fieberkurve um eine nicht-triviale Funktion handelt, zeigt dieses Beispiel auch, dass komplexe Funktionalitäten realisiert werden können.

Um die Effektivität des Ansatzes nachzuweisen müssen zwei Aspekte überprüft werden:

- Der ursprüngliche Angriff funktioniert nicht mehr und
- die Funktionalität bleibt erhalten.

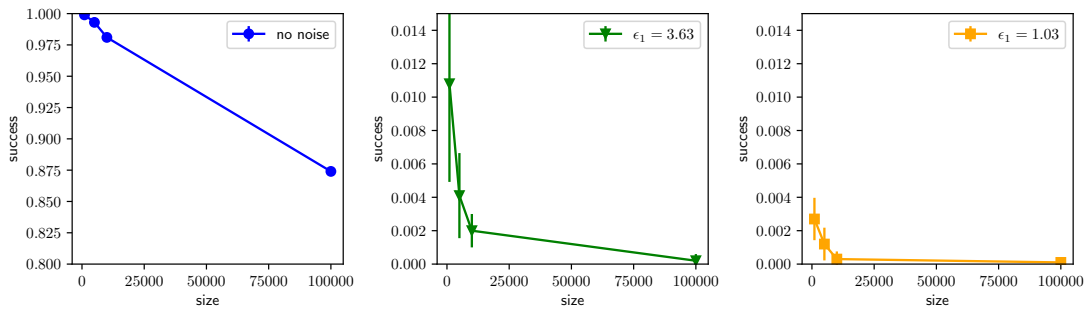
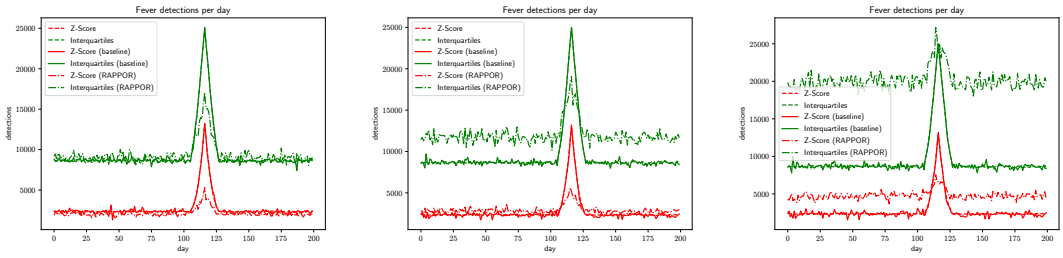


Abb. 3.4: Vergleich der Erfolgswahrscheinlichkeit des Angreifers in Bezug auf unterschiedliche Fehlerraten [16].

In Abbildung 3.4 wird die Erfolgswahrscheinlichkeit des Angreifers in Abhängigkeit der Datenbankgröße dargestellt. Dabei ist in der linken Grafik zu sehen, dass die Erfolgswahrscheinlichkeit des Angreifers bei kleinen Datenmengen nahe der 100% und bei einer Datenbank mit 100.000 Einträge noch bei über 85% liegt.

In der mittleren und rechten Grafik ist zu sehen wie sich die Einführung eines Fehlers, d. h., die Verrauschung der Daten, auf die Erfolgswahrscheinlichkeit des Angreifers auswirkt. Unabhängig von der Art des Fehlers sieht man deutlich, dass die Technik auch bei kleinen Datenbanken sehr effektiv ist und das Re-Identifikationsrisiko auf unter 2% fällt.

Was die Berechnung der Funktionalität, also der eigentlichen Fieberkurven angeht, so sind die Ergebnisse in Abbildung 3.5 dargestellt. Die Grafiken vergleichen unterschiedliche Ansätze und Fehlerraten. Die wichtigsten Aspekte der Darstellung sind folgende: Die durchgängige grüne Line stellt die eigentliche Fieberkurve ohne das Verrauschen dar und die Linie mit dem Vermerk RAPPOR stellt das Ergebnis nach dem Verrauschen dar. Die Grafiken zeigen deutlich, dass die Berechnung der Fieberkurve möglich ist.



(a) $\epsilon_1 = 1.5$ **(b)** $\epsilon_1 = 1.34$ **(c)** $\epsilon_1 = 1.03$

Abb. 3.5: Berechnungen der Fieberkurve unter Verwendung von unterschiedlichen Fehlerraten. [16]

3.3 Berechnung auf verschlüsselten Daten

In diesem Abschnitt werden verschiedene Techniken zur Berechnung auf verschlüsselten Daten vorgestellt.

3.3.1 Homomorphe Verschlüsselung

Zum besseren Verständnis wie Berechnungen auf verschlüsselten Daten durchgeführt werden können, wird zunächst die Funktionsweise einer „herkömmlichen“ Verschlüsselung erläutert. Dabei geht es nicht um die Beschreibung eines konkreten Verfahrens, sondern vielmehr um die Beschreibung der abstrakten Funktionsweise. Diese Funktionsweise wird durch Schnittstellen repräsentiert und jede Schnittstelle ermöglich eine bestimmte Operation, die von jedem Verschlüsselungsverfahren realisiert wird. Unabhängig von der Art des Verschlüsselungsverfahrens, stellt jedes Verfahren im Wesentlichen die Schlüsselerzeugung, Verschlüsselung und Entschlüsselung zur Verfügung. Da die Schlüsselerzeugung bei den folgenden Betrachtungen keine Rolle spielt, wird auf diese Operation nicht weiter eingegangen.

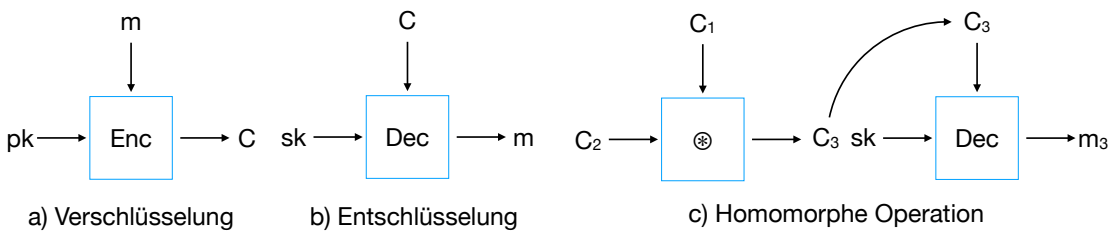


Abb. 3.6: Visualisierung einer Verschlüsselung sowie homomorpher Operationen

Ein Verschlüsselungsverfahren ist in der Kryptographie ein Verfahren, welches einen lesbaren Text, den sogenannten Klartext, in eine Form überführt, die keine Informationen über den Klartext preisgibt, das sogenannte Chiffrat. Eine Visualisierung der Ver- und Entschlüsselungsoperation ist in Abbildung 3.6 a) und b) gegeben. Der Verschlüsselungsalgorithmus Enc erhält als Eingabe eine Nachricht m und einen öffentlichen Schlüssel pk. Das Ergebnis der Berechnung ist ein Chiffrat C. Um dieses Chiffrat zu entschlüsseln, wird der Entschlüsselungsalgorithms Dec verwendet. Dieser benötigt als Eingabe den geheimen Schlüssel sk und das Chiffrat C.

Additive und Multiplikative Homomorphe Verschlüsselung

Zur Beschreibung eines homomorphen Verschlüsselungsverfahrens wird zunächst der Begriff des Homomorphismus erläutert. Dieser Begriff stammt aus der Mathematik und ist eine strukturerhaltende Abbildung zwischen zwei Strukturen. Vereinfacht gesagt bedeutet dies, dass die Ausführung einer Berechnung in einer Struktur auf eine Berechnung in einer anderen Struktur abgebildet werden kann. Diese Abbildung erhält die Struktur, d. h., die Berechnung verändert das Ergebnis nicht. Diese abstrakte Beschreibung wird durch ein anschauliches Beispiel im Folgenden dargestellt.

Der Begriff des Homomorphismus wird nun auf Verschlüsselungsverfahren übertragen. Dabei wird eine Berechnung auf dem Chiffrat durchgeführt und zwar *ohne* das Chiffrat zu entschlüsseln. Das Besondere an dieser Operation ist, dass sich die Berechnung auf den Klartext *innerhalb des Chiffrats* auswirkt und dies obwohl der Klartext bei der Berechnung nicht zur Verfügung steht. Je nach Art der Verschlüsselung wirkt sich die Verknüpfung der Chiffrate als Addition oder als Multiplikation auf die Klartexte aus. Im ersten Fall wird von einer *Additiven Homomorphen* im zweiten Fall von *Multiplikativen Homomorphen* Verschlüsselung gesprochen. Bildlich ist dies in Teil c) in Abbildung 3.6 dargestellt. In der linken Hälfe des Teils c) sind zwei Chiffrate dargestellt C_1 und C_2. Diese Chiffrate wurde auf herkömmliche Art und Weise berechnet, d. h., C_1 ist das Ergebnis der Verschlüsselung einer Nachricht m_1 mit dem öffentlichen Schlüssel pk und C_2 das Ergebnis der Verschlüsselung einer Nachricht m_2 mit dem öffentlichen Schlüssel pk. Diese beiden Chiffrate werden nun durch eine Berechnung $\circledast$ miteinander verknüpft. Folgendes Beispiel soll die Funktion dieser Operation veranschaulichen:

- Zunächst wird die Zahl 2 mit dem öffentlichen Schlüssel pk verschlüsselt: $\mathsf{C}_1 := \mathsf{Enc}(\mathsf{pk}, 2)$. Das resultierende Chiffrat wird mit C_1 bezeichnet.
- Im zweiten Schritt wird nun die Zahl 3 mit dem öffentlichen Schlüssel pk verschlüsselt $\mathsf{C}_2 := \mathsf{Enc}(\mathsf{pk}, 3)$ und wir bezeichnen mit C_2 das resultierende Chiffrat.
- Im dritten Schritt werden nun die beiden Chiffrate durch die Operation $\circledast$ miteinander verknüpft, das Ergebnis dieser Verknüpfung wird mit

  $$\mathsf{C}_3 := \mathsf{C}_1 \circledast \mathsf{C}_2$$

  bezeichnet. Handelt es sich bei dem Verschlüsselungsverfahren um ein additive homomorphes Verfahren, dann enthält das Chiffrat

  $$\mathsf{C}_3 := \mathsf{C}_1 \circledast \mathsf{C}_2 = \mathsf{Enc}(\mathsf{pk}, 2) \circledast \mathsf{Enc}(\mathsf{pk}, 3) = \mathsf{Enc}(\mathsf{pk}, 2 + 3) = \mathsf{Enc}(\mathsf{pk}, 5)$$

  den Wert 5. Ist das Verfahren multiplikativ homomorph, dann enthält

  $$\mathsf{C}_3 := \mathsf{C}_1 \circledast \mathsf{C}_2 = \mathsf{Enc}(\mathsf{pk}, 2) \circledast \mathsf{Enc}(\mathsf{pk}, 3) = \mathsf{Enc}(\mathsf{pk}, 2 * 3) = \mathsf{Enc}(\mathsf{pk}, 6)$$

In beiden Fällen ist das Ergebnis der Berechnung nur für den Besitzer des geheimen Schlüssels sk zugänglich, da nur dieser das Chiffrat C_3 entschlüsseln kann. Ein Beispiel für ein additives homomorphes Verschlüsselungsverfahren ist das so genannte ElGamal Verschlüsselungsverfahren [24], das Paillier Verschlüsselungsverfahren ist multiplikativ homomorph [25].

Vollhomomorphe Verschlüsselung

Durch die Erfindung von additiv und multiplikativen homomorphe Verschlüsselungsverfahren konnten bereits zahlreichen Anwendungen realisiert werden, welche die Klartexte von Individuen schützen und lediglich die Ergebnisse einer Berechnung preisgeben. Im Jahr 2009 stellte Gentry das erste *vollhomomorphe* Verschlüsselungsverfahren (engl.

fully homomorphic) vor [26]. Im Gegensatz zu den vorherigen Verfahren können bei dieser Art von Verschlüsselungsverfahren *beliebige* Berechnung auf dem Chiffrat ausgeführt werden. Zum besseren Verständnis der Funktionsweise werden die einzelnen Schnitte mit Hilfe von Abbildung 3.7 dargestellt und beschrieben.

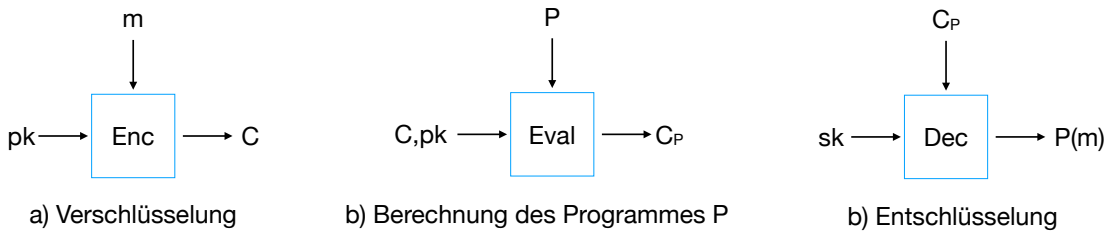


Abb. 3.7: Visualisierung der Operationen eines vollhomomorphen Verschlüsselungsverfahrens.

Im ersten Schritt Abbildung 3.7 (a) wird ein Klartext m mit dem öffentlichen Schlüssel pk verschlüsselt und das Ergebnis dieser Operation ist ein Chriffrat C. Im Teil (b) von Abbildung 3.7 wird nun die Berechnung visualisiert. Dabei handelt es sich um Programm Eval, welches als Eingabe neben einem Chiffrat C und dem öffentlichen Schlüssel pk die Beschreibung eines Programmes P als Eingabe erhält. Das Ergebnis dieser Berechnung ist ein Chiffrat $\mathsf{C_P}$, welches das Ergebnis der Berechnung enthält:

$$\mathsf{Eval}(\mathsf{pk}, C, P) = \mathsf{C_P} = \mathsf{Enc}(\mathsf{pk}, \mathsf{P}(\mathsf{m})).$$

Genau wie in den vorherigen Beispielen kann die Berechnung auf dem Chiffrat C von jedem durchgeführt werden, da keine geheimen Informationen für diese Art der Berechnung notwendig sind. Das Ergebnis kann jedoch nur von dem Besitzer des zugehörigen geheimen Schlüssels sk entschlüsselt werden. Vollhomomorphe Verschlüsselung ist ein unglaublich mächtiges Werkzeug, da es keine Beschränkungen für das Programm P gibt, d. h., sämtliche Berechnung können auf der Verschlüsselung durchgeführt werden.

Seit der ersten Entdeckung des ersten Verfahrens wurden zahlreiche neue Konstruktionen gefunden, welche die Effizienz deutlich verbessern konnten.

Beispiel

Als Beispiel nehmen wir an, dass das Chiffrat C die Verschlüsselung aller Patientendaten enthält. Auf diesen Daten sollen nun statistische Auswertungen durchgeführt werden. Beispielsweise könnte man berechnen wie viele Menschen in Berlin im Alter zwischen 18-79 an Diabetes erkrankt sind. Diese Abfrage würde als ein Programm P formuliert und dann mittels der vollhomomorphe Verschlüsselung auf C ausgewertet. Das Ergebnis der Berechnung würde nun in dem Chiffrat $\mathsf{C_P}$ gespeichert. Der Besitzer des privaten Schlüssels sk könnte nun das Ergebnis durch die Entschlüsselung von $\mathsf{C_P}$ erhalten. Dieses einfache Beispiel kann auf beliebig komplexe Szenarien erweitert werden. Soll das Ergebnis der Berechnung nur von einer authentifizierten dritten Person lesbar sein, zum Beispiel von einer Treuhandstelle, so kann das Ergebnis der Berechnung auch in einem

Chiffrat gespeichert werden, welches sich nur von dem geheimen Schlüssel der Treuhandstelle entschlüsseln lässt. Diese könnte dem Forscher nach der Entschlüsselung das Ergebnis $\mathsf{P}(\mathsf{m})$ bereitstellen.

3.3.2 Kontrollierte Berechnungen auf Verschlüsselten Daten

Die vorherigen Techniken führen Berechnungen auf verschlüsselten Daten durch, ohne dass der Betroffene die Möglichkeit hat dieser Operation zu widersprechen oder zuzustimmen. Dies ist insbesondere bei medizinischen Studien problematisch, da die Daten nach Abschluss einer Studie vernichtet werden müssen. Zu diesem Zweck wurde an dem Lehrstuhl für Angewandte Kryptographie an der FAU ein Verfahren entwickelt, welches die sensitiven Daten in einer verschlüsselten Form speichert und jede Berechnungen auf diesen Daten muss vorher autorisiert werden. Diese Autorisierung kann entweder direkt durch den Patienten oder beispielsweise durch eine Ethikkommission geschehen [27]. Im folgenden wird die Beschreibung des Systems anhand von genetischen Daten geschehen mithilfe der Darstellung in Abbildung 3.8.

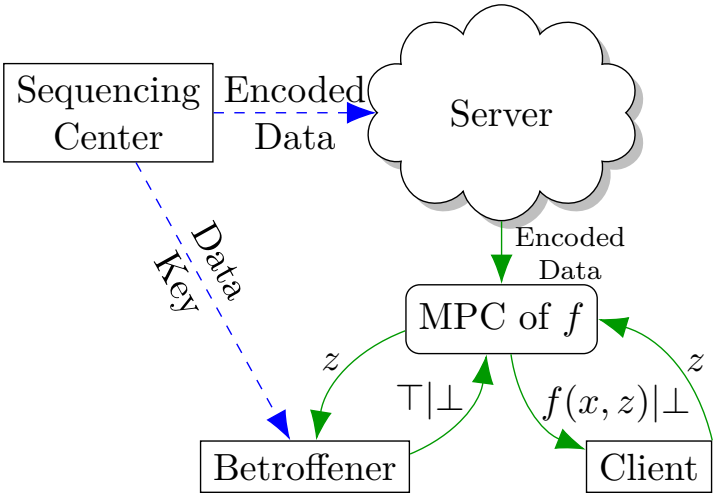


Abb. 3.8: Beschreibung des Systems [27]

In der Beschreibung des Systems gibt es vier Teilnehmer: Das Sequencing Center, der Server, der Betroffene der Daten und der Client. Im ersten Schritt wird das Genom sequenziert (Sequencing Center) und in verschlüsselter/codierter Form auf den Server geladen. Des Weiteren wird ein Schlüssel, der zur Autorisierung der Berechnung benötigt wird, an den Betroffenen ausgehändigt. Möchte nun ein Client Berechnungen auf den genetischen Daten x ausführen, so beschreibt er die Art der Berechnung z. Die Art der Berechnung wird nun an den Betroffenen (oder eine Ethikkommission) weitergeleitet (über das System zum Schutz der Privatsphäre des Patienten). Stimmt der Betroffene der Berechnung zu, in Abbildung 3.8 dargestellt durch das Symbol $\top$, so führt der Client gemeinsam mit dem Server diese Berechnung durch und erhält das Ergebnis $f(x, z)$ der Berechnung. Andernfalls, sollte die Berechnung nicht autorisiert werden, erhält der Client

nichts, dargestellt durch das Symbol $\perp$. Der Kern dieser Berechnung ist ein so genanntes „Secure Multi-Party“ (MPC) Protokoll. Diese Protokolle haben die Eigenschaft, dass die Teilnehmer in dem Protokoll nichts über die geheimen Eingaben der anderen Teilnehmer lernen. In diesem Fall bedeutet es, dass der Betroffene zwar den geheimen Schlüssel kennt und in der Ausführung verwendet, jedoch lernen weder der Server noch der Client etwas über Schlüssel.

Beispiel

Zum besseren Verständnis wird das Verfahren an folgendem fiktiven Beispiel erläutert. In diesem Beispiel sollen verschiedene Studien über Brustkrebs anhand des BRCA1 Gens durchgeführt werden.

Beginn der Studie Zu Beginn dieser Studie rekrutiert die Universität 200 Teilnehmerinnen. Die Teilnehmerinnen geben eine Blutprobe ab, welche an das Sequencing Center geschickt wird. Nach der Sequenzierung erhält die Patientin einen kryptographischen Schlüssel und die verschlüsselten Daten werden in die Cloud geladen.

Durchführung verschiedener Studien Im zweiten Schritt möchte die Universität nun verschiedene Studien auf den Daten durchführen. Zu diesem Zweck stellt die Universität eine Anfrage an das System, diese Anfrage besteht aus einer algorithmischen Beschreibung der Untersuchung und einer Erläuterung für die Teilnehmerinnen, welche an diese weitergeleitet wird. Die Teilnehmerinnen erhalten eine Nachricht auf dem Handy und erteilen gegebenenfalls ihr Einverständnis.

Berechnung der Ergebnisse Wurde das Einverständnis durch die Teilnehmerinnen erteilt, dann erhält die Universität lediglich das Ergebnis der Berechnung und damit der Studie. Die Universität erlangt keinen direkten Zugriff auf die Daten. Wurde das Einverständnis nicht erteilt, dann erhält die Universität keinerlei Information.

Herausforderung in der Realisierung

Die Herausforderung in der Realisierung eines solchen Systems liegt in der unglaublichen Datenmenge: die Speicherung des Genoms eines einzelnen Patienten benötigt ungefähr 1,5 GB. Komprimierte Formate, wie zum Beispiel das VCF Dateiformat, geben Informationen über den Patienten preis und können daher nicht angewendet werden. Neben den großen benötigen Speicherkapazitäten ist auch die Kommunikationskomplexität ein herausforderndes Problem, da zur Freigabe einer Berechnung nicht das ganze verschlüsselte Genom übertragen werden kann. Andernfalls würde folgende triviale Lösung das Problem lösen: Der Client sendet seine Anfrage an die Cloud. Die Cloud sendet das vollständige Genom zusammen mit der Anfrage an die Teilnehmer*innen. Diese führen die Berechnung auf dem Handy aus und schicken das Ergebnis der Berechnung zurück an die Cloud, die das finale Ergebnis an den Client weiterleitet. Neben offensichtlichen Sicherheitsproblemen, wie zum Beispiel dem Problem das die Cloud das Ergebnis der Berechnung lernt, stellt die Kommunikationskomplexität ein großes Problem dar. In unserer Lösung hingegen ist die Kommunikationskomplexität *unabhängig* von der Größe

des Genoms und damit sehr effizient. Die experimentelle Evaluation des Systems konnte problemlos Berechnungen auf Datensätzen von 10.000 Patienten durchführen.

Das Beispiel dieses Systems zeigt, dass selbst auf sehr großen Datenmengen komplexe Berechnungen innerhalb einer Verschlüsselung durchgeführt werden können.

3.3.3 Vor- und Nachteile dieser Ansätze

Die vorgestellten Ansätze unterscheiden sich in folgenden Aspekten:

- der Mächtigkeit der berechenbaren Funktionen,
- der Effizienz sowie
- der Notwendigkeit von Interaktion.

Generell muss man bei allen Verfahren einen Kompromiss zwischen diesen drei Punkten eingehen. Je einfacher eine Berechnung, desto effizienter kann diese auf verschlüsselten Daten realisiert werden. Je nach Anwendung ist eine interaktive Lösung wie Unterabschnitt 3.3.2 beschrieben vorzuziehen. Bei einfach statistischen Auswertungen können dedizierte Lösungen entwickelt werden.

3.4 Zusammenfassung der Ergebnisse und ein Vergleich der unterschiedlichen Ansätze und deren Qualität in Bezug auf Forschungszwecke

In diesem Kapitel wurden alternative Ansätze zur Bereitstellung von (lediglich) pseudonymisierten Datensätzen vorgestellt. Die Vielzahl an grundlegend unterschiedlichen Ansätzen zeigt, dass viel Forschung in diesem Gebiet betrieben wird und dass abhängig von der Anwendung verschiedene Techniken eingesetzt werden können, welche die Privatsphäre schützen und gleichzeitig neue Anwendungen ermöglichen. Diese Techniken nicht einzusetzen ist reine Fahrlässigkeit und sollte gesetzlich verpflichtend in die Verarbeitung von personenbezogenen Daten eingeführt werden.

Die vorgestellten Ansätze unterscheiden sich stark in verschiedenen Aspekten:

- (Beweisbare) Garantien
- Verständnis-Hürden
- Komplexität
- Einsatz kryptographischer Techniken
- Kommunikationskomplexität

Zunächst unterscheiden sich die Verfahren stark in den erreichten Sicherheitsgarantien. Die Klasse der einfachen Anonymisierungstechniken geben aus meiner Sicht die schwächsten Garantien, da relativ einfache Hintergrundinformationen das Anonymitätsset

drastisch reduzieren können. Besonders deutlich wurde dies an dem Beispiel der Analyse der Daten des Statistischen Bundesamtes, bei der wir zeigen konnten, dass das Wissen „Nutzer trägt eine Smart Watch“ das Anonymitätsset im Schnitt um 30% reduziert. Diese Art der Technik erweckt den Eindruck eines „Katz- und Mausspiels“, bei dem eine neue Technik verfeinert wird, um einen Angriff gegen die alte Technik zu mildern. Dies war in den Anfängen der Kryptographie üblich, bis sich der Begriff der beweisbaren Sicherheit etablierte und erstmals formale Modelle mit formalen Garantien aufgestellt wurden. Diese einfachen Techniken sind damit aus meiner Sicht zur Sicherung von medizinischen Daten ungeeignet, da das Hintergrundwissen des Angreifers nicht abgeschätzt werden kann und da es schwierig bis unmöglich ist eine geeignete „Schutzreserve“, die zukünftigen Risiken vorbeugt, einzuplanen (vgl. Roßnagel [3]). Des Weiteren müsste bei der Verwendung dieser Technik eine kontinuierliche Prüfung stattfinden, ob eine Verknüpfung der Daten zu einer Person in der Zwischenzeit möglich sein könnte. Auch dies erscheint im praktischen Einsatz unrealistisch und kostenintensiv.

Der große Bereich der Differential Privacy kann insofern als logische Weiterentwicklung gesehen werden, als präzise formale Modelle und formale Sicherheitsgarantien zum Tragen kommen. Des Weiteren gelten die Sicherheitsgarantien unabhängig von möglichem Hintergrundwissen und der (unbeschränkten) Rechenkapazität des Angreifers. Folglich gelten die Sicherheitsgarantien für *jedes mögliche* Hintergrundwissen, dies schließt das Hintergrundwissen ein was aktuell verfügbar ist, und auch das Hintergrundwissen was der Angreifer in der Zukunft erhält. Mit diesen Garantien geht jedoch eine Erhöhung der Komplexität einher. Sowohl das Modell, das Verständnis für die Garantien, als auch die Umsetzung sind deutlich anspruchsvoller und werden in der Regel nur von großen Technikfirmen, wie Google oder Apple umgesetzt. Dennoch sind diese Techniken für statistische Auswertungen aufgrund der verbesserten Garantien gegenüber den einfachen Anonymisierungsversuchen vorzuziehen.

Orthogonal zu dem Ansatz der Differential Privacy sind die Ansätze der Berechnung auf verschlüsselten Daten und der sicheren kontrollierten Berechnung auf verschlüsselten Daten zu sehen. Die Sicherheitsgarantien sind sicher die stärksten, da diese Techniken lediglich das Ergebnis der Berechnung preisgeben. Gleichzeitig ist die Umsetzung dieser Technik aufwendiger. Zum einen, da die verwendeten kryptographischen Techniken teilweise komplizierter sind und da man eine Lösung für das Schlüsselmanagement benötigt. Das Problem des Schlüsselmanagements stellt die Frage nach der Schlüsselverwaltung. Wer speichert die Schlüssel, wie wird der Zugriff auf die Schlüssel geregelt, was passiert mit „verlorenen“ Schlüsseln und wie werden die Schlüssel aktualisiert?

Implikationen für das Datentransparenzverfahren

In diesem Abschnitt wurden verschiedene Techniken zum Schutz der Privatsphäre vorgestellt und hinsichtlich der (beweisbaren) Garantien, den Verständnishürden, der (Kommunikations-)komplexität und dem Einsatz kryptographischer Techniken verglichen. Die Ergebnisse dieses Abschnittes zeigen, dass es eine Fülle an gut erforschten Verfahren gibt, deren Robustheit und Sicherheit teilweise seit Jahrzehnten untersucht und belegt wurden. Eine Auswahl bzw. eine Kombination dieser Verfahren sollte zwingend in

das Datentransparenzverfahren integriert werden. Denn durch diese Techniken können die sensiblen medizinischen Daten zu Forschungszwecken genutzt werden, ohne die Privatsphäre eines Einzelnen zu gefährden.

4 Sichere Speicherung der Daten

Fragestellung: Birgt die zentrale Speicherung sensibler Daten größere Risiken als eine dezentrale Speicherung? Welche Alternativen gibt es zur zentralen Speicherung von Datensätzen im Forschungsdatenzentrum und sind sie von vergleichbarer Qualität?

Dieser Abschnitt ist in folgende Teile gegliedert. In Abschnitt 4.1 wird die allgemeine Entwicklung von Cybersicherheitsangriffen besprochen und die Art von Angriffen klassifiziert. Der Wert der Daten wird in Abschnitt 4.2 abgeschätzt. Ein Vergleich zwischen einer zentralen und dezentralen Speicherung der Daten erfolgt im Abschnitt 4.3, mögliche Alternativen werden im Abschnitt 4.4 skizziert und eine Zusammenfassung der Ergebnisse erfolgt im Abschnitt 4.5.

4.1 Entwicklung von Cybersicherheitsangriffen

Zur Beantwortung der Frage nach der sicheren Speicherung der Daten wird zunächst auf die Entwicklung von Cybersicherheitsangriffen im Allgemeinen eingegangen. Dies ist insofern wichtig, da die Fragen nach der Sicherheit auch die Frage nach der Relevanz impliziert. Handelt es sich um ein rein akademisches Problem, oder ist es eine reale Bedrohung mit zunehmender Wichtigkeit? Auch die Art der Durchführung von Cyberangriffen und die ausgewählten Ziele der Angreifer sind essenziell, um zu verstehen, ob die Angriffe auf diesen Anwendungsbereich übertragbar sind oder nicht. In den folgenden zwei Abschnitten werden beide Aspekte diskutiert.

4.1.1 Entwicklung von Cybersicherheitsangriffen

In den letzten Jahren kann eine sehr starke Zunahme an Cybersicherheitsanriffen beobachtet werden [28, 29, 30, 31]. Die entstanden Kosten durch diese Angriffe wachsen seit Jahren stetig an und der Schaden wird nur für das Jahr 2021 weltweit auf 6 Billionen USD geschätzt [32]. Würde dieser Wert als Wirtschaftsleistung eines Landes zählen, so läge diese Ökonomie weltweit auf Platz 3 hinter den USA und China. Damit ist der erzielte Profit in diesem Gebiet deutlich größer, als der erwirtschaftete Gewinn aus allen illegalen Drogengeschäften weltweit [32]. Aufgrund der stetigen Digitalisierung wird ein Wachstum für diesen Bereich mit 15% pro Jahr vorhergesagt, sodass von einem wirtschaftlichen Schaden in Höhe von 10.5 Billionen USD im Jahr 2025 ausgegangen werden muss.

Diese Zahlen verdeutlichen das enorme finanzielle Interesse in diesem Bereich und wirft die Frage nach den Opfern auf. Vereinfacht gesagt sind alle Bereiche des Lebens betroffen, sämtliche Zweige der Industrie, den medizinischen Einrichtungen, dem öffentlichen Sektor und auch Forschungseinrichtungen, sie alle werden Opfer von Cyberattacken [29]. Zwar sind alle Bereiche von Cyberangriffen betroffen, jedoch lassen sich zwei Dinge aus den vorhandenen Daten ableiten: Erstens, größere Institutionen sind stärker betroffen, da mehr Daten gleichzeitig einen höheren Gewinn implizieren. Zweitens, der medizinische Bereich ist deutlich stärker von Cyberangriffen betroffen als andere Bereiche [33]. Dies liegt vermutlich an dem höheren Wert der Daten, auf den im Abschnitt 4.2 eingegangen wird.

4.1.2 Klassifikation der Angriffe

Das Ponemon Institute und IBM [29] untersuchten ebenfalls die Arten der Angriffe, welche zu dem Verlust von Daten führten. Dieser Teil der Studie gibt Aufschluss darüber, ob diese Angriffe für die betrachtete Anwendung von Relevanz sind. Im Folgenden beziehe ich mich auf die Daten in Abbildung 4.1. Folgende Angriffe sind aus meiner Sicht für

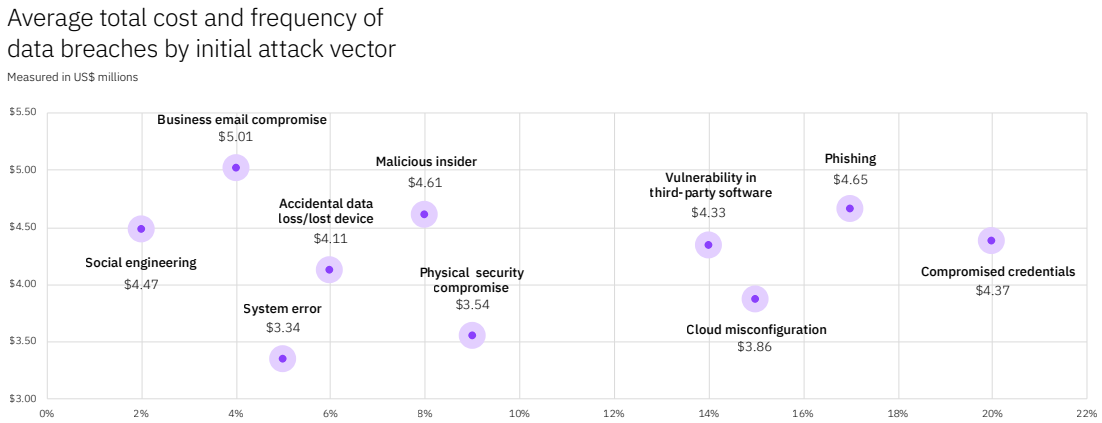


Abb. 4.1: Übersicht über die Arten der Cyberangriffe [29].

das Anwendungsgebiet von Relevanz:

Kompromittierte Anmeldeinformationen Die Studie zeigt, dass 20% aller Cyberangriffe von kompromittierten Anmeldeinformationen ausgehen. Bei dieser Art des Angriffs erlangt der Angreifer Zugang zu den Anmeldeinformationen, wie zum Beispiel dem Benutzernamen und dem Passwort, oder im Falle einer SSL Authentifikation zu dem privaten Schlüssel. Dieser Angriff ist für die vorgesehene Anwendung aus mindestens zwei Gründen sehr problematisch. Erstens, der Angriff ist äußert relevant und wird höchstwahrscheinlich auch durchgeführt werden. Zweitens, dieser Angriff ist äußerst schwer festzustellen, da eine Interaktion mit dem System über die üblichen Schnittstellen erfolgt. Dies wird auch von der Studie bestätigt, im Schnitt dauert die Detektion dieses Angriffes mit 250 Tagen am längsten. Die Behebung

dieses Angriffes dauert im Schnitt noch weitere 91 Tage, sodass die Lücke nahezu für ein Jahr besteht.

Bösartige Insider Laut der Studie entstanden 8% aller Angriffe durch bösartige Insider. Unter einem bösartigen Insider versteht man einen der derzeitigen oder ehemaligen Mitarbeiter, Auftragnehmer oder vertrauenswürdigen (Geschäfts-)Partner, der seinen autorisierten Zugriff auf sensitive Daten missbraucht. Aufgrund des enormen Wertes der hier betroffenen Gesundheitsdaten halte ich diese Form eines Angriffs für äußerst realistisch und relevant. Eine einfache konservative Abschätzung zeigt, die Patientendaten mindestens 7,3 Milliarden Euro Wert sind (siehe Abschnitt 4.2). Es dürfte schwierig werden genug geeignete Mitarbeiter*Innen zu finden, die Zugriff auf die Daten haben und bei Bestechungsgeldern im Millionenbereich nicht schwach werden.

Weitere Fälle Weitere Fälle, die in Abbildung 4.1 dargestellt sind, wie zum Beispiel versehentlicher Verlust, Systemfehler, Fehler in der Konfiguration (der Cloud), und Schwachstellen in der Software von Dritten, sind ebenfalls relevant für das hiesige Anwendungsszenario.

Zusammenfassend komme ich zu dem Schluss, dass ein Großteil der Angriffe relevant, realistisch und auf das hiesige Anwendungsszenario übertragbar sind. Im nächsten Abschnitt beschäftige ich mich mit dem Wert der Daten.

4.2 Wert der Daten

Um das Risiko der zentralen Speicherung der Daten besser beurteilen zu können, wird in diesem Abschnitt auf den Wert der Daten eingegangen. Dieser Wert ist insofern wichtig, da „wertlose“ Daten keines besonderen Schutzes bedürfen. Der Wert von Daten lässt sich auf zwei unterschiedliche Arten abschätzen: Zum einen gibt es eine Studie die seit 17 Jahren jährlich durch das Ponemon Institute und IBM durchgeführt wird [29]. In dieser Studie werden Firmen nach dem entstandenen Schaden nach einem Cyberangriff befragt. Die zweite Möglichkeit besteht in der Abschätzung des Wertes der Daten durch einen Verkauf auf dem Schwarzmarkt. Die erste Studie ist insofern von Relevanz, da sie aufzeigen kann, ob eine Bedrohung in der Praxis relevant ist. Darüber hinaus lassen sich Vorhersagen über die Zukunft ableiten. Handelt es sich um ein Problem, welchem in der Zukunft mehr Beachtung geschenkt werden muss, oder ist davon auszugehen, dass der Schutz der Daten zukünftig gewährleistet werden kann? Des Weiteren kann ein Interesse an einer bestimmten Art von Daten abgleitet werden. Treten zum Beispiel Cyberangriffe vermehrt in einem Bereich auf, so kann daraus geschlossen werden, dass der Angreifer in diesem Bereich den größten Profit erwirtschaftet.

4.2.1 Finanzieller Schaden eines Cyberangriffs

In diesem Abschnitt werden die relevanten Ergebnisse der Studie des Ponemon Instituts und IBM vorgestellt [29]. In dieser Studie werden Firmen weltweit befragt die Opfer von

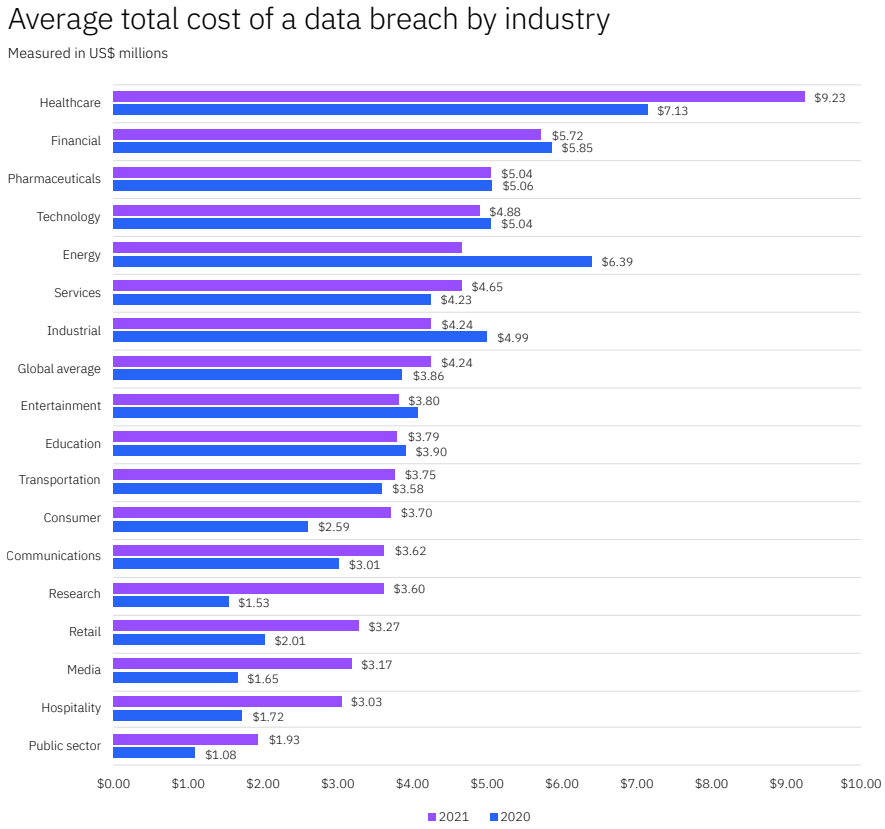


Abb. 4.2: Durchschnittliche Kosten für den Verlust von Daten durch einen Cyberangriff [29].

Cyberattacken wurden. Aus diesen Antworten errechnet die Studie dann die Kosten, die durch den Cyberangriff verursacht wurden. Des Weiteren gibt die Studie auch wichtige Einblicke in die Art der Angriffe und die betroffenen Bereiche.

Wichtige Ergebnisse der Studie

In diesem Abschnitt werden die wichtigsten Ergebnisse der Studie zur Klärung der Fragestellung zusammengefasst. Die angegebenen Werte spiegeln den *weltweiten Durchschnittswert* wider. Falls die Werte für Deutschland explizit angegeben wurden, so werden diese verwendet. Im Jahr 2021 wurden 537 Firmen aus 17 Ländern und 17 unterschiedlichen Industriebereichen befragt.

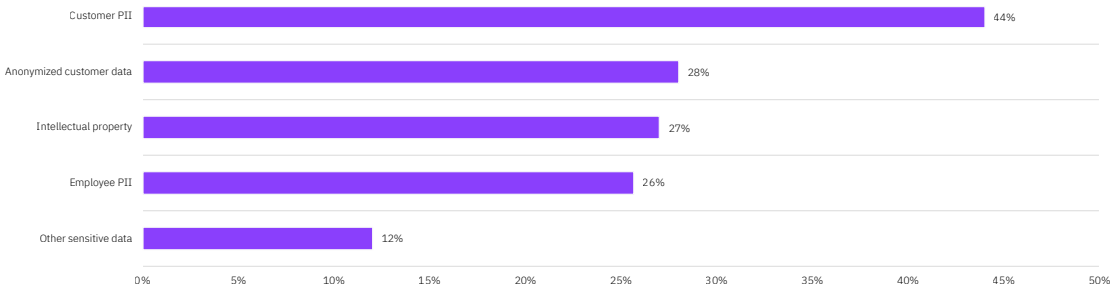
Durchschnittliche Kosten Die durchschnittlichen Kosten, die einer Firma nach einem Cyberangriff in Deutschland entstehen, werden mit 4.89 Millionen USD *pro Angriff* angegeben. Mit diesem Wert liegt Deutschland auf Platz 4 nach den USA, dem Nahen Osten und Kanada. Gemessen an den gestohlenen Daten können die entstanden Kosten *pro Datenbankeintrag* bestimmt werden. Der Durchschnittswert

pro Eintrag liegt damit bei 161 USD. Dieser Betrag ist ein Durchschnittswert über alle untersuchten Branchen. Eine Übersicht über die Kosten in Abhängigkeit der Branche ist in Abbildung 4.2 gegeben und zeigt deutlich, dass der Bereich der medizinischen Daten die höchsten Kosten verursacht.

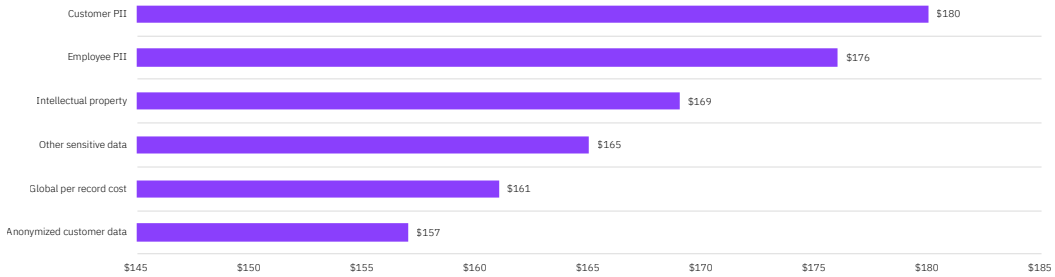
Neben der betroffenen Industrie können die Kosten ebenfalls anhand der Art des entwendeten Datentypes differenziert werden (siehe Abbildung 4.2).

Entwicklung Da diese Studie nun seit 17 Jahren durchgeführt wird, kann ein Trend abgeleitet werden. Dieser Trend zeigt deutlich, dass die Anzahl an Cyberangriffen zunimmt. Alleine im Vergleich zum vergangenem Jahr sind die entstanden Kosten um über 10 Prozent gestiegen.

Medizinische Daten Seit 11 Jahren in Folge wird der Verlust von medizinischen Daten als am kostspieligsten identifiziert. Im Vergleich zum letzten Jahr stiegen die Kosten hier um durchschnittlich 29,5%.



(a) Entwendeten Daten [29].



(b) Durchschnittliche Kosten [29].

Abb. 4.3: Übersicht über die Art der entwendeten Daten durch Cyberangriffe und die zugehörigen durchschnittlichen Kosten. PII steht für Personally Identifiable Information.

Art der entwendeten Daten Abbildung 4.3a gibt eine Übersicht über die Art der entwendeten Daten. Die Darstellung zeigt deutlich, dass die entwendeten Daten vermehrt aus identifizierenden Kundendaten, anonymisierten Kundendaten, Angestellten Daten und weiteren sensiblen Daten bestehen. Aus dieser Statistik lässt sich folgern, dass die Angreifer gezielt nach identifizierenden Daten suchen und

auch anonymisierte Daten von großem Interesse für die Angreifer sind. Aus Abbildung 4.3b wird deutlich, dass auch der Verlust von anonymisierten Daten mit sehr hohen Kosten verbunden ist. Der Verlust eines einzelnen Datenbankeintrages in diesem Fall zieht immer noch Kosten in Höhe on 157 USD im Schnitt nach sich.

Bedeutung der Datenmenge Die Studie zeigt auch, dass der Verlust von großen Datenmengen deutlich kostspieliger ist als der von kleinen Datenmengen. Insbesondere sind die durchschnittlichen Kosten für den Verlust von Datenbanken mit über 50 Millionen Einträgen im Schnitt 100x teurer als der Verlust von kleinen Datenbanken bis zu 100,000 Einträgen. Etwas weniger als die Hälfte der entstanden Kosten lassen sich in Bußgelder und in die Reparatur des Systems aufgliedern. Diese beiden Punkte sind aus meiner Sicht für den hier dargestellten Anwendungsfall, also das Datentransparenzverfahren, relevant.

4.2.2 Wert der Daten auf dem Schwarzmarkt

Ein alternativer Ansatz zur Bestimmung des Wertes der Daten kann über eine Abschätzung des Kaufpreises der Daten auf dem Schwarzmarkt erfolgen. In diesem Bereich gibt es relativ wenige Studien und auch die Art wie diese Studien erhoben wurden erschließt sich mir aus den Unterlagen nicht. Insgesamt kommen diese Studien jedoch auf sehr vergleichbare Ergebnisse und geben den Wert von medizinischen Daten auf dem Schwarzmarkt im Schnitt mit 250 USD pro Datenbankeintrag an [34, 35]. Als obere Grenzen werden Werte von 1000-2600 USD angegeben [36]. Die erzielten Preise schwanken stark je nachdem um welche Art von Daten es sich handelt und wie vollständig ein Profil ist. Als Gründe für die hohen Preise werden verschiedene Punkte aufgeführt: Beispielsweise kann die hohe Informationsdichte über einen Betroffenen dazu benutzt werden, um den persönlichen und finanziellen Ruf zu schädigen. Des Weitern können die Daten zur Erpressung genutzt werden oder um sich die Identität des Betroffenen anzueignen*. All diese Beispiele sind natürlich auch auf den hier betrachteten Anwendungsfall, die im Datentransparenzverfahren verarbeiteten Daten, übertragbar.

Zwischenfazit zum Wert der Daten

Die Ergebnisse aus den vergangen beiden Abschnitten können wie folgt zusammengefasst werden:

- Die Anzahl an Cybersicherheitsattacken nimmt stetig zu. Dies liegt aus meiner Sicht zum Teil an dem Prozess der Digitalisierung, bei dem Systeme miteinander verknüpft werden und wurden, die nicht dafür vorgesehen waren. Beispielsweise werden immer mehr Systeme in Krankenhäusern miteinander vernetzt, ohne dass die Systeme dafür konzipiert wurden. Dies führte zum Beispiel zu einem Ausfall der IT nach einem Cyberangriff im Uniklinikum Düsseldorf im September 2020. Dies ist nur eines von zahlreichen Beispielen.

*https://blog.tbconsulting.com/why-healthcare-data-is-so-valuable-on-the-black-market

- Ein weiterer Grund für die Zunahme an Cybersicherheitsattacken sind die großen Fortschritte im Bereich der Hardware. Dieser Fortschritt erlaubt die Bearbeitung großer Datenmengen durch KI Verfahren und dadurch steigt die Bedeutung und der finanzielle Wert von Daten weiter an.

- Medizinische Daten verursachen die höchsten Kosten bei Verlust und erzielen die höchsten Preise auf dem Schwarzmarkt. Diese beiden Ergebnisse zeigen, dass medizinische Daten eines besonderen Schutzes bedürfen und dass aufgrund der hohen Gewinnspannen davon ausgegangen werden muss, dass diese Art der Daten im Fokus von Cyberangriffen stehen werden.

- Alle Daten haben einen Wert, selbst anonymisierte Daten werden bei Cyberangriffen entwendet und verkauft.

- Größere Datenmengen erzeugen einen höheren Schaden und erreichen bessere Preise auf dem Schwarzmarkt.

4.3 Zentrale vs. Dezentrale Speicherung der Daten

Die hier betroffenen Daten sind aufgrund des hohen Wertes einem enormen Risiko ausgesetzt sind und zwar unabhängig von der Art der Speicherung. Dies folgt direkt aus den Ergebnissen der Studien und selbst wenn man diese Zahlen für überzogen hält, so kommt ein enormer Wert durch die große Datenmenge zustanden, was das folgende einfache Rechenbeispiel demonstrieren soll:

Zunächst wird der Gesamtwert der Daten konservativ geschätzt. In Deutschland gibt es ca. 73 Millionen gesetzlich versicherte Patienten, die ihre Daten bereitstellen müssen. Folglich besteht die Datenbank aus mindestens 73 Millionen Einträgen. Der Mittelwert des Wertes eines einzelnen Datenbankeintrags liegt bei 250 USD. Zur Vereinfachung rechnen wir nun mit einem extrem pessimistisch Wert eines Eintrages von 100 Euro. Bei 73 Millionen Datenbankeinträgen kommen wir folglich auf einen Gesamtwert der gespeicherten Daten von mindestens 7,3 Milliarden Euro. Diese Abschätzung ist sicher extrem konservativ, andere Abschätzungen gehen von einem Gesamtwert in Höhe von 150 Milliarden Euro aus [36]. An dieser Stelle sei noch erwähnt, dass diese Abschätzungen sich *auf den aktuellen Stand* beziehen. Da die Daten 30 Jahre gespeichert werden sollen, dürfte sich der Wert jährlich noch erhöhen.

Das Ziel des Angreifers ist die Maximierung seines Gewinns. Folglich wird der Angreifer einen bestimmten Betrag investieren, um Zugang zu den Daten zu erlangen. Nehmen wir nun an, dass ein Angreifer 10 Millionen Euro zur Bestechung eines Administrators investiert, so liegt die Gewinnspanne immer noch bei einem Faktor von über 700. Wie die Studien zeigen, sind diese Abschätzung eher pessimistisch, die realen Preise und auch die realen Gewinnspannen liegen sicher deutlich höher.

Des Weiteren muss man sich bewusst sein, dass der Verlust der Daten nicht revidiert werden kann. Daten die einmal entwendet wurden, verbreiten sich rasant im Internet und die Verfolgung und Löschung aller Kopien ist unmöglich.

Zentrale Speicherung der Daten

Aus meiner Sicht birgt die zentrale Speicherung der Daten deutlich größere Risiken als ein dezentraler Ansatz aus den folgenden Gründen:

Minimalitätsprinzip Die zentrale Speicherung der Daten ignoriert eines der grundlegendsten Prinzipen der IT Sicherheit, das sogenannten *Minimalitätsprinzip*. Dieses Prinzip besagt, dass nur so viel Information herausgegeben wird, wie zwingend notwendig ist. Im Falle einer Kompromittierung reduziert man durch die Anwendung dieses Prinzips den Schaden, da so wenige Informationen wie möglich preisgegeben wurden. Durch die zentrale Speicherung der Daten erreicht man das Gegenteil. Wird das System einmal kompromittiert so lernt die Angreifer sämtliche Information.

Teile und Herrsche Ein weiteres Prinzip der IT Sicherheit, welches durch die zentrale Speicherung der Daten missachtet wird, ist bekannt unter dem Begriff *Teile und Herrsche*. Dieser Begriff leitet sich aus dem Design von Algorithmen ab und folgt der grundlegenden Idee, dass das Vertrauen und die Aufgaben auf mehrer Entitäten aufgeteilt werden müssen und jede Entität herrscht über ihren Teil. Dieses Prinzip findet man beispielsweise bei jedem Bankschließfach wieder, welches zwei Schlüssel zum Öffnen des Schließfaches benötigt. Der Kunde vertraut der Bank insofern, dass diese den Inhalt eines Schließfaches sicher aufbewahrt, aber der Kunde ist sich nicht sicher, ob einer der Mitarbeiter*Innen in das Schließfach schaut. Aus diesem Grund werden zum Öffnen des Schließfaches zwei Schlüssel benötigt, von denen einer immer beim Kunden liegt. Durch die zentrale Speicherung der Daten reduziert sich das Vertrauen auf eine einzelne Entität.

Diversifizierung Der Begriff der Diversifizierung stammt ursprünglich aus der Betriebswirtschaftslehre und beschreibt eine Aufteilung des Risikos auf unterschiedliche Bereiche zur Verbesserung der Gewinnchance und Reduktion des Risikos. Gleiche Ansätze finden sich im Bereich der IT Sicherheit wieder indem versucht wird gezielt unterschiedliche Systeme zu verwenden um das Risiko eines Angriffes und dessen Auswirkungen zu minimieren. Die Verwendung unterschiedlicher Systeme hilft insofern, dass sich Angriffe oftmals gegen ein bestimmtes System und/oder eine bestimmte Software richten. So lässt sich ein Angriff auf ein Windows System in der Regel nicht auf ein Linux oder Apple-System übertragen. Durch die Zentralisierung des Systems wird das Vertrauen auf eine einzelne Konfiguration gerichtet.

Aufgrund dieser Aspekte komme ich zusammenfassend zu dem Schluss, dass die zentrale Speicherung der Daten deutlich größere Risiken birgt als eine dezentrale Lösung. Es ist aus meiner Sicht nicht nachvollziehbar, warum grundlegende Eckpfeiler der IT Sicherheit ignoriert wurde und ein extrem großes Risiko zur Speicherung der Daten eingegangen wird. Wie bereits beschrieben handelt es sich bei den Daten um besonders schützenswerten Daten die eine einen enormen Wert auf dem Schwarzmarkt erzielen.

Folglich sollte bereits bei der Konzeption der Sicherheitsarchitektur der aktuelle Stand des Wissens einfließen, dies ist bei der Art der Speicherung der Daten jedoch nicht erfolgt.

4.4 Alternative Ansätze zur Speicherung und Verarbeitung der Daten

Die Entwicklung einer Forschungsplattform, welche die Privatsphäre der Patienten schützt und gleichzeitig die Auswertung der Daten zu Forschungszwecken gewährleistet, ist ein komplexes Problem, welches einer genauen Analyse bedarf. Diese Analyse muss sowohl die Schutzziele festlegen, als auch ein Vertrauensmodell und ein Angreifermodell spezifizieren. Eine solche Analyse ist sehr umfangreich und kann nicht Bestandteil dieses Gutachtens sein. Die folgenden Ansätze stellen höchstens die Eckpfeiler eines solchen Designs dar und erheben keinen Anspruch auf Vollständigkeit. Das folgende Konzept integriert die wesentlichen Prinzipien der IT Sicherheit (siehe Abschnitt 4.3) und besteht aus folgenden drei Komponente:

Dezentrale Speicherung Verschlüsselter Daten Der größte (Mehr-)Wert der Daten entsteht durch die Verknüpfung von einzelnen Attributen. Diese Art von Verknüpfungen gilt es zu schützen und darf nur unter bestimmten Bedingungen herstellbar sein. Aus diesem Grund erfolgt eine dezentrale Speicherung der Daten. Beispielsweise könnten Quasi-Identifikatoren, wie zum Beispiel die Postleitzahl und das Geschlecht, getrennt von den Daten der Behandlung gespeichert werden. Je nach Anwendung und Art der Daten, sollten die Daten verschlüsselt gespeichert werden und die Berechnung sollte stets in der Verschlüsselung stattfinden (siehe Abschnitt 3.3), sodass am Ende lediglich das Ergebnis preisgegeben wird. Sollte der direkte Zugriff zur Bestimmung der Parameter und der Erkennung von Mustern notwendig sein, so schwebt mir ein zweistufiges System vor. Im ersten Schritt erhalten die Forscher*Innen Zugang zu einer repräsentativen Teilmenge, die mittels Techniken der Differential Privacy geschützt wurden (siehe Abschnitt 3.2). Im zweiten Schritt formulieren die Forscher dann ein Programm zur automatisierten Auswertung der Daten, welches auf den verschlüsselten Daten ausgeführt wird. Aufgrund der großen Datenmenge ist eine händische Auswertung nicht möglich und folglich kann diese auch auf verschlüsselten Daten stattfinden.

Verteilte Berechnung Die Daten sollten zu keinem Zeitpunkt direkt und im Klartext zusammengefügt werden. Stattdessen erfolgt eine verteilte Berechnung bei der die einzelnen Teilnehmer*Innen ihren Teil zu der Berechnung beitragen. Beispielsweise könnte ein Teilnehmer die (verschlüsselten) Quasi-Identifikatoren und eine zweite Teilnehmerin die (verschlüsselte) Art der Behandlung beitragen. Ein dritter Teilnehmer könnte dann das eigentliche Ergebnis der Berechnung erhalten. Dieser Ansatz stellt sicher, dass kein Teilnehmer zu einem Zeitpunkt eine komplette Sicht auf die Daten und das Ergebnis erhält.

Privacy Guthaben Innerhalb der Berechnung würde ich ein „Privacy Guthaben“ einführen. Dieses Guthaben stellt sicher, dass eine Re-Identifizierung eines Individuums

nicht möglich ist. Sollte das Guthaben unter einen bestimmten Wert fallen und damit eine potenzielle Re-Identifizierung zulassen, so würde die Berechnung kein Ergebnis ausgeben. Die Notwendigkeit eines solchen Guthabens entsteht durch wiederholte Analyse eine Datensatzes bei dem das Ergebnis der Berechnung die mögliche Menge an Personen sukzessiv reduziert oder den eingeführten Fehler im Bereich der Differential Privacy aushebelt.

4.5 Zusammenfassung der Ergebnisse

In diesem Abschnitt wurde eine Bewertung des Risikos der zentralen Speicherung der Daten vorgenommen und mit einer dezentralen Speicherung der Daten verglichen. Basierend auf den Ergebnissen komme ich zu dem Schluss, dass eine zentrale Speicherung der Daten ein deutlich größeres Risiko birgt, da lediglich eine zentrale Instanz angegriffen werden müsste. Des Weiteren wurde deutlich, dass die zentrale Speicherung wesentliche Eckpfeiler der IT Sicherheit ignoriert, wie zum Beispiel das Minimalitätsprinzip und die Aufteilung des Vertrauens auf mehrere Teilnehmer*Innen. Im Abschnitt 4.4 wurden verschiedene Ideen alternativer Realisierungen skizziert. Dieser Abschnitt zeigt zwei Dinge: Erstens wurde bei der Konzipierung der Architektur der Stand der Wissenschaft nicht berücksichtigt und bewusst eine leicht angreifbare Architektur ausgewählt. Zweiten gibt es bereits heute genug Alternativen, die zu einer sicheren Realisierung führen würden. Technisch gesehen bietet die zentrale Speicherung der Daten aus meiner Sicht keinen vertretbaren Mehrwert, da die gleichen Berechnungen auch dezentral ausgeführt werden könnten (siehe hierzu auch das Beispiel der verteilten Datenaggregation mittels *Private Set Intersection* in Abschnitt 5.3). Folglich bin ich der Meinung, dass die sensiblen Daten der Bürger nicht entsprechend dem aktuellen Stand von Wissenschaft und Technik gesichert wurden.

5 Datensammelstelle

Ist die Datensammelstelle technisch erforderlich? Könnten nicht die Krankenkassen selbst die Daten prüfen und parallel ein Lieferpseudonym an die Vertrauensstelle schicken?

Dieses Kapitel beschäftigt sich mit der technischen Notwendigkeit der Datensammelstelle und gliedert sich in folgende Abschnitte. Der Datentransfer wird in Abschnitt 5.1 beschrieben und intuitive Sicherheitsgedanken dieses Designs werden im Abschnitt 5.2 diskutiert, gefolgt von einer Evaluation im Abschnitt 5.3. Ein alternativer Ansatz, der ohne die Datensammelstelle auskommt, wird im Abschnitt 5.4 skizziert. Ein Vergleich zwischen dem aktuellen Ansatz und dem alternativen Vorschlag erfolgt im Unterabschnitt 5.4.2.

5.1 Darstellung des Datentransfers

In diesem Abschnitt wird der Datentransfer zwischen den einzelnen Teilnehmer*Innen mit Hilfe von Abbildung 5.1 beschrieben. Im ersten Schritt senden die Krankenkassen

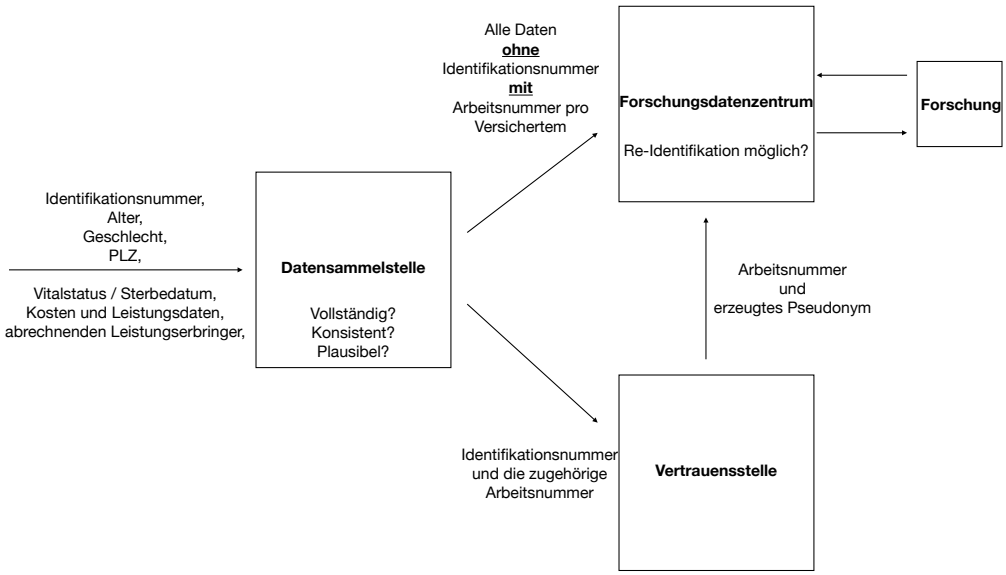


Abb. 5.1: Darstellung des Datentransfers nach § 303b und § 303C SGB V.

nach § 303 b Abs. 1 - 5 SGB V unter anderem folgende Informationen an die Datensam-

melstelle (siehe auch § 3 DaTraV):

- Identifikationsnummer
- Alter
- Geschlecht
- PLZ
- Vitalstatus / Sterbedatum
- Kosten und Leistungsdaten
- abrechnenden Leistungserbringer

Diese führt dann alle Daten zusammen und überprüft die Daten auf Vollständigkeit, Konsistenz und Plausibilität (§ 303 b Abs.2 DaTraV). Nach Abschluss dieser Prüfungen übermittelt die Sammelstelle folgende Daten:

Forschungsdatenzentrum Alle Daten bis auf die Identifikationsnummer/Versichertenkennzeichen aus § 303 b Abs.1 SGB V werden an das Forschungszentrum geschickt. Des Weiteren werden die Einzeldatensätze mit einer Arbeitsnummer versehen.

Vertrauensstelle Die Vertrauensstelle erhält alle Versichertenkennzeichen und Arbeitsnummern. Sie erzeugt Einweg-Pseudonyme und schickt diese zusammen mit den Arbeitsnummern an das Forschungsdatenzentrum. Nach dem Transfer der Daten werden die Daten gelöscht (§ 303 c Abs. 1-3 SGB V).

Nach § 303 d Abs. 1-3 SGB V stellt das Forschungsdatenzentrum die Daten zu Forschungszwecken zur Verfügung und übernimmt verschiedene Aufgaben, wie zum Beispiel die der Qualitätssicherung, Prüfung der Datennutzung und der Bestimmung des Re-Identifikationsrisikos (siehe Kapitel 6).

5.2 Intuitive Sicherheitsgedanken für das Design des Datentransfers

Der grundlegende Entwurf des Datentransfers scheint auf folgenden Überlegungen zu beruhen:

- Eine Überprüfung der Daten auf Vollständigkeit, Konsistenz und Plausibilität kann nur erfolgen, wenn die Stelle ein vollständiges Bild über alle Daten hat.
- Um den Zusammenhang zwischen den Patienten und Behandlungen zu verschleiern wird den einzelnen Positionen eine Arbeitsnummer gegeben und das Versichertenkennzeichen entfernt. Diese „pseudonymisierten Daten“ werden dann an das Forschungsdatenzentrum übertragen.

- Damit die Daten wieder zugeordnet werden können ohne auf den Patienten schließen zu können, bedarf es einer pseudonymen Zuordnung. Dies wird durch die Vertrauensstelle realisiert indem die Datensammelstelle die Versichertenkennzeichen und Arbeitsnummern erhält und ein Pseudonym berechnet.

Aus diesen Punkten kann nun folgendes zugrundeliegendes Vertrauensmodell abgeleitet werden:

- Die Datensammelstelle muss als vollkommen vertrauenswürdig angesehen werden, da sie Zugriff auf sämtliche Daten erhält.
- Dem Forschungsdatenzentrum wird insofern vertraut, dass es die Daten sicher speichert, nicht modifiziert und die Arbeitsnummern durch die entsprechenden Pseudonyme ersetzt. Des Weiteren führt das Zentrum keine Versuche zur Re-Identifikation durch und kommt seinen Aufgaben in Bezug auf die Durchführung von Forschungsvorhaben nach.
- Der Vertrauensstelle wird in dieser Konstellation das wenigste Vertrauen beigemessen, da diese lediglich die Versichertenkennzeichen und Arbeitsnummern erhält und keinen direkten Zugriff auf die Daten. Folglich hat die Vertrauensstelle keinen direkten Zugriff auf sensible Daten und die (alleinige) Kompromittierung würde lediglich zufällige Werte preisgeben.

5.3 Evaluation des Vertrauensmodells des Datentransfers

Das beschriebene Design des Datentransfers ist aus den folgenden Gründen aus meiner Sicht nicht nachvollziehbar:

- Die Notwendigkeit der Datensammelstelle wird damit begründet, dass nur so eine Überprüfung auf Vollständigkeit, Konsistenz und Plausibilität möglich sei. Diesen Punkt kann ich nicht nachvollziehen, da aus meiner Sicht diese Überprüfungen auch bei den Krankenkassen direkt bzw. in einem verteilten Algorithmus zwischen den Krankenkassen durchgeführt werden könnte:
  - Einfache Probleme, wie zum Beispiel leere Tabellen, könnten die Krankenkassen selber bereinigen.
  - Komplexe Aufgaben könnten zwischen den Krankenkassen ausgeführt werden ohne die Datenschutzbestimmungen zu verletzen. Wechselte ein Patient zum Beispiel die Krankenkasse, so wäre es die Aufgaben der Datensammelstelle die Daten von beiden Krankenkassen zusammenzufassen, um ein konsistentes Bild zu erhalten. Selbst diese Aufgabe kann kryptographisch gelöst werden indem beide Krankenkassen ein so genanntes *Private Set Intersection* Protokoll ausführen. In diesem Protokoll wird die Schnittmenge zweier Mengen berechnet *ohne*, dass die einzelnen Teilnehmer*Innen Informationen über die geheimen Patientendaten des anderen Teilnehmers lernen. Das Ergebnis dieser Berechnung ist die Schnittmenge der Patienten, die in beiden Versicherungen

aufgeführt werden. Dieses Ergebnis könnte innerhalb des kryptographischen Protokolls anonymisiert werden und bei einer dritten Partei gespeichert werden.

  - Weitere Anforderungen, wie zum Beispiel das Finden von Duplikaten kann ebenfalls kryptographisch leicht abgebildet werden.

- Besteht man jedoch auf die Notwendigkeit einer Datensammelstelle, dann erschließt sich mir nicht die Notwendigkeit der Vertrauensstelle. Wie im Abschnitt 5.1 beschrieben, wird die Datensammelstelle als vollkommen vertrauenswürdig angesehen, da diese Zugang zu sämtlichen Daten erhält. Folglich kann die Datensammelstelle direkt die Berechnung der Pseudonyme vornehmen, da diese die Daten und die Zuordnung bereits kennt.

Abschließend sei noch zu erwähnen, dass die Formulierung des Gesetzes den Eindruck erweckt als reiche die Entfernung des Versichertenkennzeichens und des Namens zur Herstellung der Anonymität aus. Dies ist nicht der Fall, da alle Angriffe aus Kapitel 2 hier natürlich zum Tragen kommen. Die vorgestellten Angriffe funktionieren auf Daten, die weder ein Versichertenkennzeichen noch einen Namen tragen. Folglich wurden durch die Entfernung dieser Daten keinerlei Hürden geschaffen, die eine Re-Identifikation erschweren. Im Gegenteil, durch die Integration von (unverrauschten) feingranularen Daten, wie der Medikation, wurde ein sehr großer Datenraum geschaffen, der eine Re-Identifikation noch vereinfacht. Eine Zuordnung der Daten würde dann wie im Unterabschnitt 2.4.2 beschrieben, erfolgen.

Zusammenfassend komme ich zu dem Schluss, dass die Konzeption der Datensammelstelle den Stand der Forschung der letzten zwei Jahrzehnte außer Acht lässt und damit unnötige Sicherheitsrisiken eingeht, die zu irreversiblen Datenverlusten führen könnten.

5.4 Alternativer Ansatz für den Datentransfer

Generell sollte die Modellierung eines solchen Systems dem „Security-and-Privacy by Design“ folgen und in einem formalen Sicherheitsmodell evaluiert werden. In solch einem Modell werden die Sicherheitseigenschaften formal beschrieben und es wird nachgewiesen, dass diese Eigenschaften innerhalb des Modells erreicht werden. Eine solche Modellierung ist nicht Teil dieses Gutachtens und es wird kein Anspruch auf Vollständigkeit erhoben.

5.4.1 Alternativer Ansatz: Die Vertrauensstelle als Hüter der Integrität

Im ersten Ansatz dient die Vertrauensstelle als eine Art Hüter über die Integrität der Daten und es wird keine Datensammelstelle benötigt. Zur Vereinfachung wird davon ausgegangen, dass die Daten bereits bereinigt wurden, d.h., es befinden sich keine Duplikate oder leeren Einträge in den Datenbanken. Außerdem wird vereinfachend angenommen, dass die Datenbanken disjunkt sind, d.h., kein Patient befindet sich in zwei Datenbanken (eine mögliche Realisierung ist in Abschnitt 5.3 beschrieben). Die einzelnen Schritte werden mithilfe von Abbildung 5.2 beschrieben und umfassen folgende Punkte:

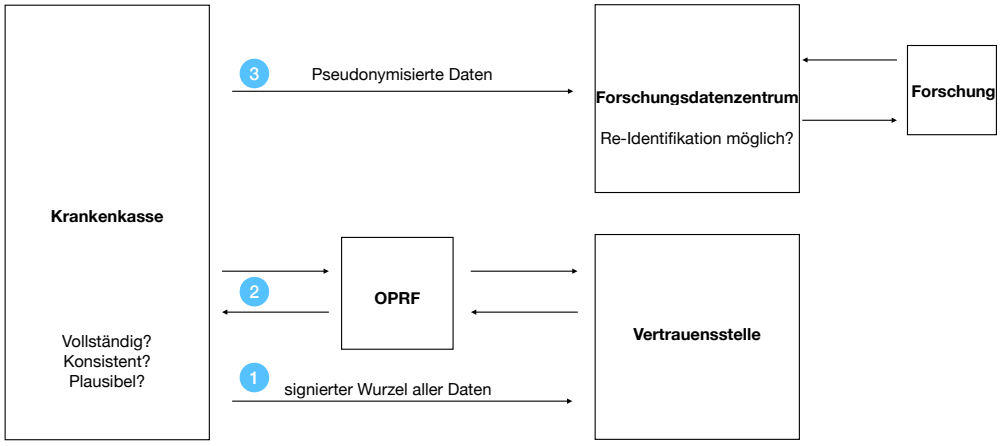


Abb. 5.2: Alternativer Ansatz zur Stärkung der Privatsphäre.

- Im ersten Schritt führt die Krankenkasse alle notwendigen Prüfungen durch. Wenn die Überprüfung abgeschlossen ist, dann legt sich die Krankenkasse auf alle Daten fest. Dies dient dazu, dass potenzielle Modifikationen im Nachhinein festgestellt werden können. Dazu berechnet die Krankenkasse einen Hashbaum* über alle Daten. Eine solche Datenstruktur stellt sicher, dass die Korrektheit einzelner Einträge sehr effizient überprüft werden kann *ohne* alle Daten erneut zu berechnen. Die Wurzel dieses Hashbaumes signiert die Krankenkasse und schickt diesen Wert an die Vertrauensstelle. Dieses Vorgehen erfüllt mehrere Dinge: Erstens, die Daten welche von der Krankenkasse übertragen wurden, können im Nachhinein nicht verändert werden, da diese Veränderung auch einen neuen Wurzelwert impliziert. Zweitens, die Vertrauensstelle lernt nichts über die eigentlichen Daten, da der Wurzelwert diese versteckt. Drittens, können in einem Audit-Verfahren einzelne Einträge effizient überprüft werden.

- Im zweiten Schritt erzeugt die Krankenkasse zusammen mit der Vertrauensstelle ein Pseudonym für die Daten. Dies ist durch die OPRF Box dargestellt, wobei OPRF für ein kryptographisches Verfahren namens *Oblivious Pseudorandom Function* steht. Das Besondere an diesem Verfahren ist, dass die Krankenkasse das Pseudonym nicht alleine erstellen kann, sondern nur mithilfe der Vertrauensstelle. Gleichzeitig kennt die Vertrauensstelle jedoch nicht die Ausgabe der Berechnung. Konkret würde die Krankenkasse eine zufällige ID wählen und das OPRF Verfahren mit der Vertrauensstelle ausführen. Am Ende des Verfahrens erhält die Krankenkasse einen pseudozufälligen Wert p. Das resultierende Pseudonym besteht nun aus der zufälligen ID und dem erhaltenen pseudozufälligen Wert p.

- Die Krankenkasse schickt die pseudonymisierten Daten zusammen mit der zufälli-

*Ein Hashbaum ist ein kryptographisches Verfahren, welches Daten in den Blättern eines binären Baumes speichert. Diese Datenstruktur stellt sicher, dass jegliche Modifikation dieser Daten sehr effizient festgestellt werden kann.

gen ID und dem pseudozufälligen Wert p im dritten Schritt an das Forschungsdatenzentrum. Das Forschungsdatenzentrum kann bei Bedarf das Pseudonym durch Interaktion mit der Vertrauensstelle überprüfen, indem es das OPRF Verfahren zusammen mit der Vertrauensstelle ausführt, dabei die zufällige ID verwendet und am Ende überprüft, ob sie den Wert p erhält. Dank der Eigenschaften der OPRF kann die Vertrauensstelle nicht feststellen, welcher Wert überprüft wurde.

5.4.2 Zusammenfassung und Vergleich der zwei Ansätze

In diesem Abschnitt wurde die Notwendigkeit der zentralen Datensammelstelle untersucht und ein alternativer Ansatz vorgestellt. Der Ansatz aus Abschnitt 5.4 bietet aus meiner Sicht folgende Vorteile im Vergleich zu der aktuellen Lösung:

- Durch den Wegfall der Datensammelstelle wird die kritischste Komponente, der sogenannte *Single Point of Failure*, entfernt. Bei dem aktuellen Design werden sich die Angriffe auf die Datensammelstelle konzentrieren, da dort alle Daten zusammengeführt werden. Der Verzicht auf diese zentrale vertrauenswürdige Einheit ist damit deutlich sicherer.

- Die Konstruktion ist konzeptionell deutlich einfacher, da nicht der Umweg über Arbeitsnummern gegangen werden muss. Stattdessen werden die Pseudonyme aufseiten der Krankenkassen mithilfe der Vertrauensstelle berechnet.

- Im Gegensatz zu der aktuellen Lösung sind alle Schritte durch kryptographische Verfahren verifizierbar. Falls es also zu Widersprüchen oder Fehlern kommt, kann dies problemlos nachvollzogen werden.

Zum Abschluss möchte ich betonen, dass dieses Design meiner Intuition folgt, es wurde kein formales Sicherheitsmodell aufgestellt und die Sicherheit wurde nicht formal bewiesen. Des Weiteren ist die Frage nach einer sicheren Übertragung und Aggregation der Daten unabhängig von der Frage nach der sicheren Speicherung und Verarbeitung der Daten.

6 Bewertung des Re-Identifikationsrisiko

Kann das Forschungsdatenzentrum das spezifische Re-Identifikationsrisiko eines Datensatzes bewerten, wie es § 303d Abs. 1 Nr. 5 SGB V verlangt?

6.1 Erhobene Daten

Zur Beantwortung der Fragen werden im ersten Schritt die Art der erhobenen Daten betrachtet. Im Groben können die Daten in zwei Klassen aufgeteilt werden: soziodemografische und medizinische Daten. Zu den soziodemographischen Daten, dargestellt in Tabelle 6.1, zählen das Geburtsjahr, Geschlecht und Postleitzahl. Die medizinischen Daten schließen zum Beispiel die Art der Behandlung, Befunde der Zahnärzte und Abgabe von Arzneimitteln ein.

Geburtsjahr	Geschlecht	Postleitzahl der Wohnortes	Sterbedatum	Betriebsnummer der Krankenkasse	Versichertenstatus	...

Tab. 6.1: Erhobene Daten nach § 303b Abs. 1 Satz 1 Nr. 1 bis 3 SGB V. Eine genauere Ausführung kann § 3 DaTraV entnommen werden.

6.2 Re-Identifikationsrisiko basierend auf soziodemographischen Daten

Im Unterabschnitt 2.4.2 wurden Re-Identifikationsangriffe auf die Datenspende App des RKIs beschrieben. Die hier verarbeiteten Daten sind vergleichbar insofern, dass das Geburtsjahr, Geschlecht und die Postleitzahl erhoben werden. Im Gegensatz zu den Daten des RKIs werden die Daten unverfälscht übertragen, und nicht wie im Falle des RKIs als generalisierte Daten. Die präzisen Daten vereinfachen die Re-Identifikation, da die Anonymitätssets relativ klein sein dürften. Um das Re-Identifikationsrisiko nur für die Verarbeitung von soziodemographischen Daten bestimmen zu können, müsste das Forschungsdatenzentrum zwei Dinge hinzuziehen:

- Die Anfragen und Ergebnisse aus der Vergangenheit müssen einbezogen werden.
- Es werden die Hintergrundinformationen, welche die anfragende Person hat, benötigt.

Der erste Punkt, also die Einbeziehung vorheriger Anfragen, ist insofern wichtig, da wiederholte Fragen die Anonymitätsmenge stark einschränken können. Zum Beispiel könnte sich eine Studie mit der Anzahl von Fehlgeburten nach Städten beschäftigen. In einer zweiten Studie geht es nun um die Anzahl von Frauen im Alter zwischen 20 - 30 in Deutschland nach Städten. Beide Studien erscheinen auf den ersten Blick nachvollziehbar. Kombiniert man jedoch die Ergebnisse beider Studien mit der Tatsache, dass die Wahrscheinlichkeit einer Schwangerschaft in diesem Altersabschnitt bei ca. 80% liegt, so kann man eine gute Vorhersage treffen, welche Frau eine Fehlgeburt hatte. In einigen Postleitzahlbereichen dürfte dies sogar eindeutig sein.

Um vergangene Anfragen einzubeziehen, könnte ein möglicher Ansatz in der Speicherung und Verarbeitung von früheren Anfragen liegen. Zwar erscheint dieser Ansatz auf den ersten Blick vielversprechend, er lässt sich jedoch nicht so ohne weiteres umsetzen: Erstens, alleine die Speicherung der vergangenen Fragen gibt sensible Informationen preis. Zweitens, es kann passieren, dass die Abschätzung, ob eine Anfrage problematisch ist oder nicht, nicht effizient berechenbar ist, da mit jeder neuen Anfrage der Raum an Möglichkeiten exponentiell wächst.

Der zweite Punkt, also die Frage nach dem Hintergrundwissen, wird auch an dem letzten Beispiel deutlich. Die Entscheidung, ob eine Anfrage ein Re-Identifikationsrisiko in sich birgt oder nicht, hängt sehr stark von dem Hintergrundwissen ab. Teilweise erscheint das Hintergrundwissen harmlos, wie zum Beispiel die Frage, ob ein Nutzer eine Smartwatch besitzt. In der Tat reduziert alleine dieses Wissen die Anonymitätsmenge um ca. 30% [16]. Erschwerend kommt hinzu, dass das Gebiet noch einen hohen Forschungsbedarf hat. Die Frage welche Hintergrundinformationen bei welchen Fragestellungen zur Re-Identifikation eines Individuums führt, wurde in vielen Gebieten noch nicht erforscht.

Zusammenfassend komme ich zum Schluss, dass die Auswertung soziodemografischer Daten ein hohes Re-Identifikationsrisiko birgt. Die Abschätzung, ob eine bestimmte Studie zu der Re-Identifikation eines Individuums genutzt werden kann, halte ich zum jetzigen Zeitpunkt nicht für realisierbar. Zum einen da es noch viele offenen Forschungsfragen gibt und da das Hintergrundwissen, was zu einer Bewertung nötig wäre, nicht abschätzbar ist. Zum anderen, da der Abgleich mit vorherigen Anfragen nicht effizient berechenbar ist.

6.3 Re-Identifikationsrisiko basierend auf medizinischen Daten

Die zweite Klasse von erhobenen Daten schließen medizinische Behandlungen ein. Beispielsweise wird hier, unter anderem, Beginn und Ende einer Behandlung, Art der Inanspruchnahme, der Erkrankungs- und Leistungsbereich, sowie die Diagnose erfasst. Aus meiner Sicht ist die Abschätzung, ob ein Re-Identifikationsrisiko vorliegt, in diesem Bereich ungleich schwieriger zu beantworten. Diese Einschätzung möchte ich durch zwei Beispiele verdeutlichen. Erstens, existiert meines Wissens nach keine Forschung, die sich mit der Möglichkeit einer Re-Identifikation basierend auf medizinischen Daten befasst. Die Arbeit über den Netflix Preis zeigte auf der theoretischen Ebene, dass eine Re-Identifikation immer dann einfach möglich ist, wenn es viele Möglichkeiten gibt, die

dünn besetzt sind [12]. Vereinfacht sieht man dies am Beispiel von der Dosierung von Medikamenten. Es gibt sehr viele Medikamente und für jedes Medikament gibt es eine bestimmte Dosierung. Es dürfte schwierig werden viele Patienten zu finden, die exakt die gleichen Medikamente in der gleichen Dosierung bekommen. Somit stellt die Dosierung der Medikamente einen eindeutigen Identifikationswert dar.

Zweitens, ist erneut das Problem des Hintergrundwissens betroffen, welches ich mit folgendem Beispiel verdeutlichen möchte. In diesem Beispiel gehen wir davon aus, dass der Angreifer medizinische Bilder, wie zum Beispiel Röntgenaufnahmen, erhalten hat. Diese Bilder können aus einer Studie stammen, oder wurden zu Forschungszwecken anonymisiert publiziert. Als konkretes Beispiel betrachten wir im Folgenden eines der größten öffentlich zugänglichen Datensätze von Röntgenbildern des Brustkorbs. Das National Institutes of Health (NIV) publizierte über 100.000 Brustströntgenbilder von über 30.000 Patienten, die für Forschungszwecke frei zugänglich gemacht wurden [37]. In Tabelle 6.2 ist eine verkürzte Form der Klassifikation dargestellt (weitere Merkmale wurden aus Platzgründen weggelassen). Neben dieser Kategorisierung sind auch die zugehörigen

Bild ID	Nachverfolgung	Diagnose	Patienten ID	Alter	Geschlecht
00000001_000.png	0	Kardiomegalie	1	57	M
00000001_001.png	1	Kardiomegalie, Emphysem	1	58	M
00000001_002.png	2	Kardiomegalie, Erguss	1	58	M
00000002_000.png	0	—	2	80	M
00000003_000.png	0	Hernie	3	74	W

Tab. 6.2: Beispieldaten aus dem Datensatz des NIH [37].

Röntgenaufnahmen frei zugänglich, Beispiele sind in Abbildung 6.1 gegeben. Die Darstellung zeigt, dass verschiedene Aufnahmen zu einer bestimmten Person über die Patienten ID zugeordnet werden können. Durch die Angabe des Altes wird auch festgehalten in welchem Jahr die Aufnahme entstand und wie sich das Krankheitsbild des Patienten entwickelt hat. Aus den Datensätzen wird ebenfalls deutlich, dass acht unterschiedliche Krankheiten in den Datensätzen dargestellt werden. Da es keine direkte Referenz auf einen Patienten gibt und lediglich das Alter sowie das Geschlecht publiziert wurde, erscheint eine Zuordnung der Bilder zu einem bestimmten Patienten nicht möglich. Im Juni 2021 veröffentlichen Packhäuser et al. eine Arbeit, die sich mit der Anonymisierung dieser Röntgenaufnahmen beschäftigt [37]. Die Autoren konnten zeigen, dass eine eindeutige Zuordnung der Bilder zu einer Person mit einer Wahrscheinlichkeit von über 95% möglich ist. Mit anderen Worten, falls ein Angreifer ein Röntgenbild von einer Person besitzt, so kann dieser mit über 95% Wahrscheinlichkeit feststellen, ob weitere Bilder in der (anonymisierten) Datenbank zu der Person gehören. Dies ist selbst dann möglich, wenn 10 Jahre zwischen den Bildern liegt. Um die Daten zuzuordnen, wendeten die Autoren Techniken aus dem Bereich des Deep Learnings an.

Die jüngsten Ergebnisse von Packhäuser et al. verdeutlichen, dass medizinische Daten, auch ohne die Angabe eines Namens, oftmals einzigartig sind. Die Einzigartigkeit eines Menschen spiegelt sich nicht nur in dem äußeren Aussehen wider, sondern ebenfalls in weniger offensichtlichen Merkmalen, wie den Röntgenbildern.

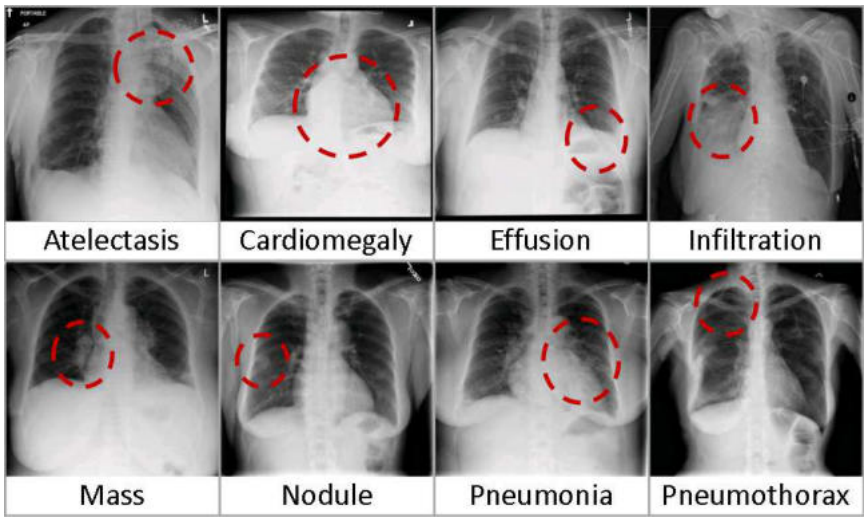


Abb. 6.1: Beispielbilder des NIH [37].

Natürlich handelt es sich bei dem Beispiel um medizinische Daten, die nicht aus Deutschland stammen. Dennoch sieht man an der Tabelle sehr schön, dass verschiedene Diagnosen über die Jahre für eine bestimmte Person gemacht wurden. Sollten diese Bilder jedoch aus Deutschland stammen, so könnte man dank der Zuordnung der Bilder, mittels der Diagnosen, eine Verknüpfung zu den Daten in der Datenbank herstellen*. Dieses Beispiel verdeutlich, dass das Forschungsdatenzentrum unmöglich eine Aussage über das Hintergrundwissen machen kann. Mit anderen Worten: Solange das Forschungsdatenzentrum nicht weiß, welches Hintergrundwissen zu einer Re-Identifikation genutzt werden kann, und ob die Forschenden dieses Hintergrundwissen besitzen, kann sie auch das Re-Identifikationsrisiko nicht einschätzen.

6.3.1 Zusammenfassung der Ergebnisse

Ziel dieses Abschnittes war die Beantwortung der Fragestellung, ob das Forschungsdatenzentrum das spezifische Re-Identifikationsrisiko eines Datensatzes bewerten kann. Basierend auf den Ergebnissen aus dem Abschnitt 6.2 und Abschnitt 6.3 komme ich zu dem Schluss, dass das Forschungsdatenzentrum das spezifische Re-Identifikationsrisiko nicht abschätzen kann. Zu diesem Schluss komme ich aufgrund der folgenden Tatsachen:

- Es mangelt in diesem Bereich an Forschungsergebnissen und Erfahrung, sodass aktuell nicht abschätzbar ist, welche Informationen eine Re-Identifikation zu welchem Grad verringern.

*Nach § 363 SBG V, kann der Versicherte die Daten aus der Patientenakte dem Forschungsdatenzentrum spenden. Solche Datenspenden können sich natürlich negativ auf die Anonymitätsmenge der Patienten auswirken, die keine Daten gespendet haben.

- Die Abschätzung des Hintergrundwissens eines Angreifers ist praktisch nicht umsetzbar. Dieses Hintergrundwissen ist für eine belastbare Abschätzung jedoch elementar.
- Außerdem fehlt insbesondere im Bereich der Medizin jegliche Erfahrung, welche Daten zur Re-Identifikation benutzt werden können.
- Des Weiteren kann die Einbeziehung aller Anfragen aus der Vergangenheit zur Bewertung des Re-Identifikationsrisikos sensitive Informationen preisgeben. Auch scheint die Menge der notwendigen Vergleiche mit den vorherigen Anfragen exponentiell anzusteigen, so dass diese Vergleiche praktisch nicht berechenbar sind.

Literaturverzeichnis

[1] S. Bretthauer and I. Spiecker gen. Döhmann, "Das Digitale-Versorgung-Gesetz als Einfallstor für eine Neujustierung von einstweiligem Rechtsschutz vor dem BVerfG und der Eingriffsqualität bei Datenverwendungen," *JuristenZeitung*, vol. 75, no. 20, pp. 990–996, 2020. Besprechungsaufsätze.

[2] S. Holst, B. Schütze, and G. Spyra, "Arbeitshilfe zur pseudonymisierung/anonymisierung." https://gesundheitsdatenschutz.org/download/Pseudonymisierung-Anonymisierung.pdf. Zugriff: 2021-07-30.

[3] A. Roßnagel, "Pseudonymisierung personenbezogener Daten," in *Zeitschrift für Datenschutz*, vol. 243, 2018.

[4] B. für politische Bildung, "Bevölkerung nach Altersgruppen und Geschlecht." https://www.bpb.de/nachschlagen/zahlen-und-fakten/soziale-situation-in-deutschland/61538/altersgruppen. Zugriff: 2021-07-30.

[5] P. Berrand, F. Knörr, and D. Schröder, "Pandemic Privacy — Breaking and Repairing the Corona App of the Robert-Koch-Insitute." Unter Begutachtung, 2021.

[6] I. Herscovici and A. Hörburger, "Alzheimerdemenz Vererbbar." https://demenz-portal.at/aktuelles/alzheimerdemenz-vererbbar/. Zugriff: 2021-11-24.

[7] L. Bothe, "Erblicher Brustkrebs – Wenn der Krebs in den Genen liegt." https://www.krebsgesellschaft.de/onko-internetportal/basis-informationen-krebs/basis-informationen-krebs-allgemeine-informationen/erblicher-brustkrebs-wenn-der-k.html. Zugriff: 2021-11-24.

[8] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006* (U. Dayal, K. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y. Kim, eds.), pp. 139–150, ACM, 2006.

[9] Y. Tao, H. Chen, X. Xiao, S. Zhou, and D. Zhang, "Angel: Enhancing the utility of generalization for privacy preserving publication," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 7, pp. 1073–1087, 2009.

[10] A. S. M. T. Hasan, Q. Jiang, H. Chen, and S. Wang, "A new approach to privacy-preserving multiple independent data publishing," *Applied Sciences*, vol. 8, no. 5, 2018.

[11] I. k. Netflix, "Trainingsdaten des netflixpreises." https://www.kaggle.com/netflix-inc/netflix-prize-data. Zugriff: 2021-08-01.

[12] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy*, pp. 111–125, IEEE Computer Society Press, May 2008.

[13] R. K. Institut, "Die Corona Datenspende App." https://corona-datenspende.de. Zugriff: 2021-08-05.

[14] R. K. Institut, "Die corona datenspende app – blog." https://corona-datenspende.de/science/. Zugriff: 2021-08-05.

[15] R. K. Institut, "Berechnung der fieberkurve." https://cybersecurityventures.com/annual-cybercrime-report-2020/. Zugriff: 2021-08-05.

[16] P. Berrang, F. Knörr, and D. Schröder, "Pandemic privacy — baking and repairing the corona app of the robert-koch-insitute." Manuskript —- Unter Begutachtung.

[17] Heise.de, "Daten-Leak bei Autovermieter Buchbinder: 3 Millionen Kundendaten offen im Netz." https://heise.de/-4643015, 2021. Zugriff: 2021-10-20.

[18] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," tech. rep., 1998.

[19] T. Jäschke, S. Rochow, H. Tewes, A. Vogel, H. Mertes, J. Reiter, and O. Methner, "Füur immer Anonym: Wie kann De-anonymisierung verhindert werden?." hhttps://www.abida.de/sites/default/files/ABIDA%20Gutachten%20F\protect\unhbox\voidb@x\bgroup\U@D1ex{\setbox\z@\hbox{\char127}\dimen@-.45ex\advance\dimen@\ht\z@}\accent127\fontdimen5\font\U@DU\egroupR%20IMMER%20ANONYM.pdf, 2019. Zugriff: 2021-10-20.

[20] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy beyond k-Anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, p. 3, Mar. 2007.

[21] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, 2007.

[22] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[23] U. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067, Nov. 2014. arXiv: 1407.6981.

[24] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," in *CRYPTO'84* (G. R. Blakley and D. Chaum, eds.), vol. 196 of *LNCS*, pp. 10–18, Springer, Heidelberg, Aug. 1984.

[25] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *EUROCRYPT'99* (J. Stern, ed.), vol. 1592 of *LNCS*, pp. 223–238, Springer, Heidelberg, May 1999.

[26] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *41st ACM STOC* (M. Mitzenmacher, ed.), pp. 169–178, ACM Press, May / June 2009.

[27] D. Deuber, C. Egger, K. Fech, G. Malavolta, D. Schröder, S. A. K. Thyagarajan, F. Battke, and C. Durand, "My genome belongs to me: Controlling third party computation on genomic data," *PoPETs*, vol. 2019, pp. 108–132, Jan. 2019.

[28] C. B. Forbes, "Alarming Cybersecurity Stats: What You Need To Know For 2021." https://www.forbes.com/sites/chuckbrooks/2021/03/02/alarming-cybersecurity-stats-------what-you-need-to-know-for-2021/?sh=4fcf0c1458d3, 2021. Zugriff: 2021-10-11.

[29] I. Security and P. Institute, "Cost of a Data Breach Report 2021." https://https://www.ibm.com/security/data-breach, 2021. Zugriff: 2021-10-01.

[30] C. for Strategic International Studies, "Significant Cyber Incidents Since 2016." https://csis-website-prod.s3.amazonaws.com/s3fs-public/211022_Significant_Cyber_Incidents.pdf?aEdoMUixpyx5OpU4dNevDfNSFfKraUgT, 2021. Zugriff: 2021-10-11.

[31] Wikipedia, "List of data breaches." https://en.wikipedia.org/wiki/List_of_data_breaches, 2021. Zugriff: 2021-10-01.

[32] S. M. Cybercrime Magazine, "Global cybercrime damages predicted to reach $6 trillion annually by 2021." https://cybersecurityventures.com/annual-cybercrime-report-2020/. Zugriff: 2021-10-10.

[33] S. M. Cybercrime Magazine, "Healthcare Industry To Spend $125 Billion On Cybersecurity From 2020 To 2025." https://cybersecurityventures.com/healthcare-industry-to-spend-125-billion-on-cybersecurity-from-2020-to-2025/. Zugriff: 2021-10-20.

[34] B. S. Experian, "Here's How Much Your Personal Information Is Selling for on the Dark Web." https://www.experian.com/blogs/ask-experian/heres-how-much-your-personal-information-is-selling-for-on-the-dark-web/, 2017. Zugriff: 2021-10-10.

[35] Trustwave, "The Value of Data." https://www.infopoint-security.de/media/TrustwaveValue_of_Data_Report_Final_PDF.pdf, 2017. Zugriff: 2021-10-10.

[36] T. Maus, "Die bomben ticken," *c't*, vol. 26, pp. 166–171, 2019.

[37] K. Packhäuser, S. Gündel, N. Münster, C. Syben, V. Christlein, and A. Maier, "Is medical chest x-ray data anonymous?," *CoRR*, vol. abs/2103.08562, 2021.