Datenschutzgerechte Forschungsschnittstelle für medizinische Daten

Bachelorarbeit

KIT – Karlsruher Institut für Technologie
Fraunhofer IOSB – Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung

Moritz Leitner

12. Oktober 2020

Verantwortlicher Betreuer: Prof. Dr.-Ing. habil. Jürgen Beyerer
Betreuender Mitarbeiter: Arno Appenzeller, M. Sc.

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die Arbeit selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissenschaftlicher Praxis in der gültigen Fassung beachtet habe.

Karlsruhe, den 12. Oktober 2020

(Moritz Leitner)

Zusammenfassung

Die Digitalisierung im Gesundheitswesen schreitet voran: Nach dem Patientendaten-Schutz-Gesetz (PDSG) müssen die Krankenkassen ihren Versicherten spätestens ab dem 1. Januar 2021 eine elektronische Patientenakte (ePA) anbieten, in der auf Wunsch beispielsweise Diagnosen, Therapiemaßnahmen oder Medikationspläne gespeichert werden. Darüber hinaus haben Versicherte ab 2023 die Möglichkeit einer Datenspende, sie können also Daten ihrer ePA der medizinischen Forschung zur Verfügung stellen. Durch die Auswertung solcher Real-World-Daten könnten Nebenwirkungen von Medikamenten in Zukunft schneller entdeckt werden.

Einer Datennutzung steht allerdings die besondere Schutzwürdigkeit von personenbezogenen Gesundheitsdaten entgegen, deren missbräuchliche Verwendung zu einer Stigmatisierung oder Diskriminierung von Betroffenen führen kann. Um den Zielkonflikt zwischen Datenschutz und Forschungsdatennutzung bestmöglich zu lösen, wurden unterschiedliche Methoden zum Schutz der Privatsphäre entwickelt, die in der Literatur gemeinhin als *Privacy-Enhancing Technologies* (PETs) bezeichnet werden.

Diese Arbeit bietet einerseits einen Überblick über den aktuellen Stand von E-Health in Deutschland. Anderseits werden die wichtigsten PETs erörtert. Dies umfasst insbesondere Anonymitätsmaße, wie k-Anonymity, ℓ-Diversity, t-Closeness, δ-Presence und *Differential Privacy* (DP), und homomorphe Verschlüsselung. Abschließend werden die vorgestellten PETs hinsichtlich ihrer Eignung für medizinische Daten untersucht. Hierfür wurde im Rahmen dieser Arbeit eine prototypische Forschungsschnittstelle namens *PRIvacy cOmpliant Research Interface* (PRIORI) entwickelt, die zur Anonymisierung und statistischen Auswertung von Datensätzen auf die Open-Source-Lösungen *ARX* und *OpenDP* setzt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AES Advanced Encryption Standard. 8

API Application Programming Interface. 39, 40, 47, 49, 53

ARXaaS ARX as a Service. 47, 51

BDSG Bundesdatenschutzgesetz. 10

BMG Bundesministerium für Gesundheit. 3

CPU Central Processing Unit. 49

CSV Comma-Separated Values. 41, 46, 50, 56

DP Differential Privacy. 2, 12, 13, 17, 19, 28–31, 33, 38, 46, 47, 50, 52, 53, 55, 56

DSGVO Datenschutz-Grundverordnung. 1, 9, 10, 13, 15, 16

DVG Digitale-Versorgung-Gesetz. 4

E-Health-Gesetz Gesetz für sichere digitale Kommunikation und Anwendungen im Gesundheitswesen. 4

eGA Elektronische Gesundheitsakte. 5

eGK Elektronische Gesundheitskarte. 4, 5, 7

eHBA Elektronischer Heilberufsausweis. 5

EHR Electronic Health Record. 3

EMD Earth Mover Distance. 25, 26

eMP Elektronischer Medikationsplan. 4, 6

ePA Elektronische Patientenakte. 1, 3–8, 55

FHIR Fast Healthcare Interoperability Resources. 5

GiB Gibibyte. 49

GMG Gesetz zur Modernisierung der gesetzlichen Krankenversicherung. 4

HL7 Health Level 7. 5

HMAC Keyed-Hash Message Authentication Code. 16

HPC High Performance Computing. 34

HSM Hardware Security Module. 8

HTTP Hypertext Transfer Protocol. 38, 40–46, 49, 51

ICD International Classification of Diseases. 5

IfSG Infektionsschutzgesetz. 11

IHE Integrating the Healthcare Enterprise. 7

IPsec Internet Protocol Security. 6

JSON JavaScript Object Notation. 5, 39–46, 65–73

KIM Kommunikation im Medizinwesen. 4

KIS Krankenhausinformationssystem. 4

KVNR Krankenversichertennummer. 8, 17, 18

LDP Local Differential Privacy. 29, 30, 33

LPDDR Low-Power Double Data Rate. 49

MBO-Ä Musterberufsordnung der Ärzte. 9

MNIST Modified National Institute of Standards and Technology. 12

NFDM Notfalldatenmanagement. 4, 5

OASIS Organization for the Advancement of Structured Information Standards. 8

ORM Object-Relational Mapping. 47

PAPAYA Platform for Privacy Preserving Data Analytics. 13

PDSG Patientendaten-Schutz-Gesetz, siehe [BT20]. 4, 6

PET Privacy-Enhancing Technology. 1, 2, 12, 15, 17, 39, 45, 49, 53, 55, 56

PKI Public-Key-Infrastruktur. 6

PPDM Privacy-Preserving Data Mining. 17

PPDP Privacy-Preserving Data Publishing. 16

PRIORI PRIvacy cOmpliant Research Interface. 37–39, 42, 44, 47–50, 52, 53, 55, 56

PVS Praxisverwaltungssystem. 4

RAM Random-Access Memory. 49

REST Representational State Transfer. 5, 12, 38, 47, 53, 55, 56

SDM Standard-Datenschutzmodell. 11

SGB Sozialgesetzbuch. 11

SGB V Fünftes Buch Sozialgesetzbuch. 4–8

SGD Schlüsselgenerierungsdienst. 8

SMC-B Elektronischer Praxisausweis. 5

SNOMED CT Systemized Nomenclature of Medicine - Clinical Terms. 5

StGB Strafgesetzbuch. 9

TI Telematikinfrastruktur. 4–6

TTP Trusted Third Party. 15

UML Unified Modeling Language. 47

URI Uniform Resource Identifier. 38, 40, 41, 45

VPN Virtual Private Network. 6

VSDM Versichertenstammdatenmanagement. 4

XACML eXtensible Access Control Markup Language. 8

XDS Cross-Enterprise Document Sharing. 7

XML Extensible Markup Language. 5

1. Einleitung

Unter den Begriff der Gesundheitsdaten fallen alle Informationen, die im Zusammenhang mit dem physischen oder psychischen Zustand eines Menschen stehen. Die Datenschutz-Grundverordnung (DSGVO) stellt besonders hohe Anforderungen an die Verarbeitung dieser hochsensiblen Informationen, da bei einer missbräuchlichen Verwendung mit einer Stigmatisierung oder Diskriminierung der betroffenen Personen gerechnet werden muss. Man möge sich allein die Nachteile bei der Jobsuche oder dem Abschluss einer Krankenversicherung vor Augen führen, die mit einer Offenlegung von Details über eine Erkrankung verbunden sind [BEE18, S. 50 f.].

Gleichzeitig besteht auch vonseiten der Politik ein wachsendes Interesse an der Nutzung der Möglichkeiten, die sich durch eine elektronische Erfassung und Auswertung von Gesundheitsdaten bieten. So sind die Krankenkassen ab dem 1. Januar 2021 verpflichtet, ihren Versicherten eine elektronische Patientenakte (ePA) anzubieten. Mit dieser Maßnahme sollen langfristig Kosten gespart werden, indem überflüssige Mehrfachuntersuchungen vermieden werden oder die Medikation bei mehreren behandelnden Ärzten besser abgestimmt wird. Außerdem bilden Gesundheitsdaten das Fundament der medizinischen Forschung, weshalb Versicherte ab 2023 ihre Daten für Forschungszwecke spenden können [BK20]. Nicht zuletzt prognostizieren Beratungsunternehmen unter dem Schlagwort *Life Sciences 4.0* große Umbrüche in der Gesundheitsbranche durch datengetriebene Geschäftsmodelle und der damit einhergehenden Neuausrichtung von *Sick Care* hin zu *Health Care* [EY18].

Um einen Ausgleich zwischen den Chancen und Risiken zu erzielen, die sich insbesondere bei der Nutzung von Gesundheitsdaten für medizinische Forschungszwecke ergeben, hat man in der Vergangenheit auf die Pseudonymisierung oder das Entfernen eindeutiger Identifikationsmerkmale gesetzt. Zwischenzeitlich gelang allerdings in unzähligen Fällen eine Re-Identifizierung einzelner Individuen in anonymisiert geglaubten Datensätzen. Auch eine aktuelle Studie kommt beispielsweise zu dem Ergebnis, dass bereits die vier Attribute Postleitzahl, Geburtsdatum, Geschlecht und Anzahl der Kinder ausreichen, um über 80 Prozent der Bewohner von Massachusetts zu identifizieren [RHM19].

Aus diesem Grund wurden in den letzten Jahren zahlreiche Methoden zum Schutz der Privatsphäre entwickelt, die in der Literatur gemeinhin als *Privacy-Enhancing Technologies* (PETs)

bezeichnet werden. Bekannte PETs sind Anonymitätsmaße, wie *k*-Anonymity und *Differential Privacy* (DP), und homomorphe Verschlüsselung. Sie ermöglichen bei richtiger Anwendung eine datenschutzgerechte Forschungsdatennutzung, indem sie zum Beispiel durch eine Verallgemeinerung der Attributwerte eine Re-Identifizierung erschweren oder die Ergebnisse von Datenbankanfragen verrauschen.

1.1. Zielsetzung

Das Ziel dieser Arbeit ist es, zunächst einen Überblick über E-Health in Deutschland und einschlägige Datenschutznormen zu bieten. Anschließend sollen im Rahmen einer umfassenden Literaturrecherche die wichtigsten PETs identifiziert und erörtert werden. Außerdem sollen die vorgestellten PETs hinsichtlich ihrer Eignung für medizinische Daten untersucht werden. In diesem Zusammenhang ist insbesondere die Performanz und die Wahl geeigneter Parameter zu evaluieren. Hierzu soll im Rahmen dieser Arbeit eine prototypische Forschungsschnittstelle entwickelt werden, die mithilfe von Open-Source-Software eine Anonymisierung und statistische Auswertung von Datensätzen erlaubt.

1.2. Struktur der Arbeit

Die vorliegende Arbeit gliedert sich in fünf Teile. In Kapitel 2 wird der aktuelle Stand von E-Health in Deutschland zusammengefasst, wobei auch die datenschutzrechtlichen Implikationen für die Forschung mit Gesundheitsdaten auf Basis bundes- und europarechtlicher Vorgaben aufgezeigt werden. In Kapitel 3 werden die bedeutendsten PETs vorgestellt, die bei der Literaturrecherche identifiziert wurden. Kapitel 4 ist den Ausführungen zum Entwurf und der Implementierung der prototypischen Forschungsschnittstelle gewidmet, welche anschließend in Kapitel 5 evaluiert wird. Das Fazit und ein Ausblick auf offene Fragen in Kapitel 6 beschließen die Arbeit.

2. Analyse

Dieses Kapitel legt den aktuellen Stand von E-Health in Deutschland dar und zeigt relevante datenschutzrechtliche Aspekte sowie verwandte Arbeiten auf.

2.1. E-Health in Deutschland

Auch wenn sich die vielfältigen Definitionen des Begriffs „E-Health“ in der Literatur im Detail unterscheiden, so ist bei allen doch der Einsatz von elektronischer Datenverarbeitung ein zentraler Aspekt. Stellvertretend sei daher die Begriffsbestimmung des Bundesministeriums für Gesundheit (BMG) genannt:

> „Unter E-Health fasst man Anwendungen zusammen, die für die Behandlung und Betreuung von Patientinnen und Patienten die Möglichkeiten nutzen, die moderne Informations- und Kommunikationstechnologien (IKT) bieten. E-Health ist ein Oberbegriff für ein breites Spektrum von IKT-gestützten Anwendungen, in denen Informationen elektronisch verarbeitet, über sichere Datenverbindungen ausgetauscht und Behandlungs- und Betreuungsprozesse von Patientinnen und Patienten unterstützt werden können […].“[1]

Im folgenden Abschnitt werden sowohl die technischen Grundlagen als auch der geschichtliche Hintergrund der in Einführung befindlichen elektronischen Patientenakte (ePA), engl. *Electronic Health Record* (EHR), behandelt, von der man sich allen voran eine Verbesserung der Versorgungsqualität sowie Kosteneinsparungen durch Vermeidung von Mehrfachuntersuchungen oder Medikationsfehlern erwartet [FK16, S. 187].

2.1.1. Geschichte

Der Lipobay-Skandal im Jahr 2001 gilt als Ausgangspunkt für die Digitalisierung des deutschen Gesundheitswesens vonseiten der Politik: Eine Wechselwirkung des Cholesterinsenkers Lipobay mit einem weiteren Medikament, das ebenfalls zur Blutfettsenkung verordnet wurde,

1 `https://www.bundesgesundheitsministerium.de/service/begriffe-von-a-z/e/e-health.html` (besucht am 26. 05. 2020)

zog weltweit 52 Todesfälle durch eine Zersetzung von Muskelgewebe nach sich, worauf der Hersteller Bayer das Medikament vom Markt nehmen musste [FP01; Sch16, S. 34].

Derartige Vorkommnisse sollten in Zukunft durch computergestützte Auswertung von Medikationsdaten verhindert werden. 2003 wurde daher das Gesetz zur Modernisierung der gesetzlichen Krankenversicherung (GMG) verabschiedet, welches die Einführung der elektronischen Gesundheitskarte (eGK) und der dafür erforderlichen „Informations-, Kommunikations- und Sicherheitsinfrastruktur“, heute bekannt als Telematikinfrastruktur (TI), zum Beginn des Jahres 2006 vorsah (§ 291a Fünftes Buch Sozialgesetzbuch (SGB V) i. d. F. v. 14.11.2003). Die mit der Umsetzung betraute Selbstverwaltung der Ärzte und Krankenkassen gründete zu diesem Zweck die gematik – Gesellschaft für Telematikanwendungen der Gesundheitskarte mbH (heute gematik GmbH) [AH16, S. 98].

Negative Erfahrungen in ersten Feldtests und Konflikte zwischen den Gesellschaftern der gematik führten schlussendlich zu einer verzögerten Einführung der eGK am 1. Januar 2015 [FK16, S. 155–159; MPE17, S. 139–145]. Das Ende 2015 beschlossene Gesetz für sichere digitale Kommunikation und Anwendungen im Gesundheitswesen (E-Health-Gesetz) sah erstmals nicht nur Fristen für den flächendeckenden Anschluss der Arztpraxen an die TI vor, sondern auch Anreize und Sanktionen in Form von Honorarkürzungen. Ärzte und Psychotherapeuten müssen laut E-Health-Gesetz seit dem 1. Juli 2019 das Versichertenstammdatenmanagement (VSDM) als erste Anwendung in der TI durchführen, bei welchem die auf der eGK gespeicherten Versichertendaten beim Einlesen in der Praxis überprüft und gegebenenfalls aktualisiert werden [Krü15; JS20, S. 94–102]. Ab dem zweiten Halbjahr 2020 können darüber hinaus E-Arztbriefe über die TI-Anwendung Kommunikation im Medizinwesen (KIM) übermittelt werden, auch der elektronische Medikationsplan (eMP) und das Notfalldatenmanagement (NFDM) stehen laut gematik dann zur Nutzung bereit [Krü20; Ärz20].

Der Zeitplan für weitere Anwendungen ist im Digitale-Versorgung-Gesetz (DVG) und im Patientendaten-Schutz-Gesetz (PDSG), dessen Entwurf am 1. April 2020 vom Bundeskabinett beschlossenen wurde, geregelt: So soll 2021 die freiwillige, versichertengeführte elektronische Patientenakte (ePA) eingeführt werden, 2022 soll das E-Rezept für verschreibungspflichtige Arzneimittel folgen. Außerdem sollen Krankenhäuser, Apotheken und fakultativ auch Pflegeeinrichtungen an die TI angebunden werden [Wey20; BK20; BT20].

2.1.2. Datenbasis

In den Primärsystemen im Gesundheitswesen, wie Praxisverwaltungssystemen (PVS) oder Krankenhausinformationssystemen (KIS), fallen verschiedenste Daten an. Das können beispielsweise Befunde, Bilddateien oder auch Abrechnungsziffern sein. Auch Versicherte selbst

erfassen in zunehmendem Maße Gesundheitswerte mit Smartphones oder Wearables („Quantified Self“) und speichern die Informationen in einer elektronischen Gesundheitsakte (eGA)[2] ab [BEE18, S. 3–7; Deu18].

Damit die Daten zwischen, aber auch innerhalb der Primärsysteme ausgetauscht werden können, ist Interoperabilität eine Grundvoraussetzung. Die beteiligten Systeme müssen also zur Kommunikation eine einheitliche, standardisierte Syntax und Semantik verwenden. Im Gesundheitsbereich haben sich auf der syntaktischen Interoperabilitätsebene, auf der Nachrichtenformat und Schnittstelle zu spezifizieren sind, die Standards der Organisation *Health Level 7* (HL7) etabliert, im Besonderen die Versionen 2 und 3 sowie *Fast Healthcare Interoperability Resources* (FHIR). FHIR setzt als neuester Standard auf eine zeitgemäße *Representational State Transfer* (REST)-Schnittstelle, als Datenaustauschformat kann *JavaScript Object Notation* (JSON) oder *Extensible Markup Language* (XML) eingesetzt werden [ML17, S. 636–646, 669–674].

Zur Erreichung semantischer Interoperabilität muss das auszutauschende Wissen einheitlich repräsentiert werden. Dies geschieht auf Basis von Klassifikationen und Terminologien, wobei letztere ein niedrigeres Abstraktionsniveau aufweisen. Klassifikationen wie *International Classification of Diseases* (ICD) werden daher eher für Abrechnungszwecke benötigt, während Terminologien wie *Systemized Nomenclature of Medicine - Clinical Terms* (SNOMED CT) bei elektronischen Aktensystemen Anwendung finden [FK16, S. 25–44; ML17, S. 674–681].

2.1.3. Telematikinfrastruktur

Wie Abbildung 2.1 veranschaulicht, lässt sich die TI in die zentrale und dezentrale TI-Plattform-Zone sowie die Provider-Zone unterteilen. Die dezentrale TI-Plattform-Zone umfasst einerseits in Form von elektronischem Heilberufsausweis (eHBA), elektronischem Praxisausweis (SMC-B) oder eGK die Smartcards der TI-Teilnehmer, andererseits die Konnektoren und Kartenterminals mit entsprechenden Gerätekarten, welche unter anderem eine transportverschlüsselte Verbindung zwischen den Hardwarekomponenten in dieser Zone ermöglichen. Die Smartcards authentisieren die Teilnehmer gegenüber der TI und erlauben zusätzlich eine direkte gegenseitige Card-to-Card-Authentisierung, weshalb die Kartenbeantragungs- und Herausgabeprozesse sicherheitskritisch sind [gem19d]. Nachdem der Chaos Computer Club erhebliche Sicherheitsmängel bei den Ausgabeverfahren nachweisen konnte, wurden die zugrunde liegenden Prozesse überarbeitet [May20]. Der Konnektor steuert die Kartenterminals im Leistungserbringernetz und bietet dem Primärsystem des Leistungserbringers Schnittstellen für die TI-Anwendungen an. Diese Schnittstellen sind im Konnektor als Fachmodule realisiert, NFDM-

2 Bei der eGA nach § 68 SGB V handelt es sich um eine Übergangslösung bis zur Einführung der ePA. Leistungserbringer sind nicht verpflichtet, Daten in eine eGA wie „Vivy“ oder „TK-Safe“ einzustellen.

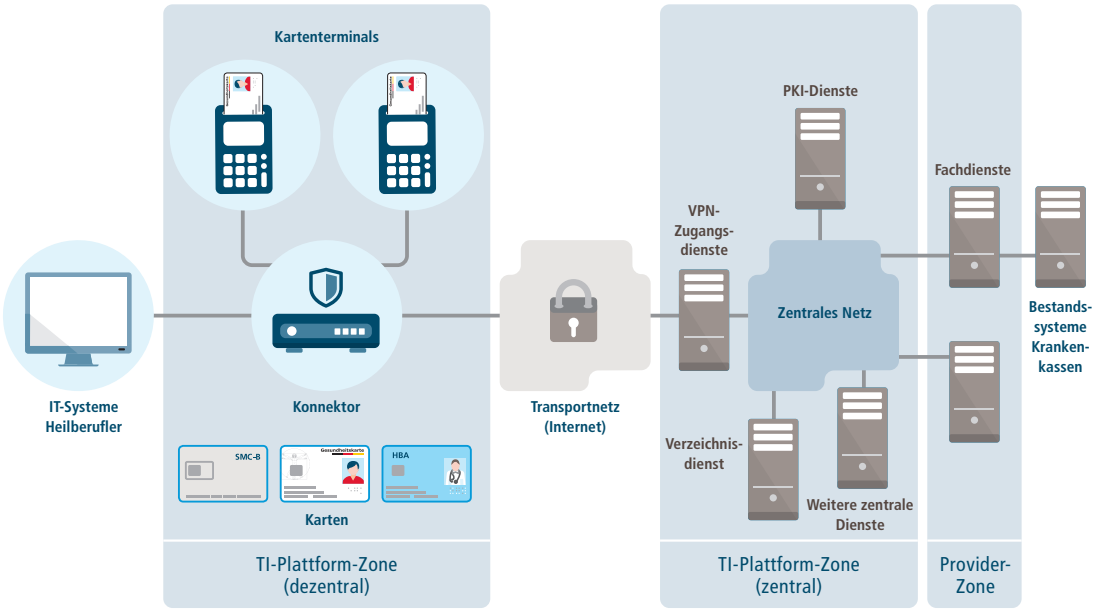

Abbildung 2.1.: Architektur der TI [gem19d]

und ePA-Fachmodul greifen hierbei auf die jeweiligen Fachdienste der Provider-Zone zurück [gem19d].

Die zentrale TI-Plattform-Zone stellt die von den Anwendungen benötigten Dienste bereit. Hierzu gehört insbesondere die Public-Key-Infrastruktur (PKI) der TI mit dem zugehörigen Verzeichnisdienst, über welche die Authentizität der Akteure sichergestellt wird. Die Verbindung zwischen den Konnektoren und dem *Virtual Private Network* (VPN)-Konzentrator, also zwischen dezentraler und zentraler TI-Plattform-Zone, erfolgt durch *Internet Protocol Security* (IPsec)-Tunnel [gem19a].

2.1.4. Elektronische Patientenakte

Mit dem im Gesetzgebungsprozess befindlichen PDSG soll der rechtliche Rahmen für die stufenweise Einführung der ePA ab dem 1. Januar 2021 festgesetzt werden, insbesondere durch Änderungen und Ergänzungen des SGB V, die im weiteren Verlauf durch SGB V-E gekennzeichnet sind. So werden die möglichen Inhalte der ePA durch § 341 SGB V-E spezifiziert: (i) Untersuchungs- und behandlungsbezogene medizinische Informationen (Befunde und Diagnosen sowie Kopien von eMP, Notfalldaten, E-Arztbriefen etc.), (ii) vom Versicherten bereitgestellte Gesundheitsdaten, (iii) der Nachweis zahnärztlicher Vorsorgeuntersuchungen (elektronisches Zahn-Bonusheft), (iv) Daten, die der Versicherte der Krankenkasse zur Verfü-

gung stellt oder von ihr zur Verfügung gestellt bekommt, (v) ein elektronischer Mutterpass, (vi) die elektronische Impfdokumentation, (vii) ein elektronisches Untersuchungsheft für Kinder, (viii) pflegerische Versorgungsdokumente, (ix) elektronische Verordnungen sowie (x) elektronische Arbeitsunfähigkeitsbescheinigungen [BT20].

In der ersten Umsetzungsstufe nach § 342 SGB V-E können zunächst nur die in (i) und (ii) genannten medizinischen Dokumente gespeichert werden. Des Weiteren ist eine differenzierte Rechtevergabe erst mit der zweiten Umsetzungsstufe ab 2022 möglich, vorher kann ein zugriffsberechtigter Leistungserbringer alle Datensätze einsehen [BK20].

Da die medizinischen Dokumente wegen begrenztem Speicherplatz nicht direkt auf der eGK abgespeichert werden können, werden die Akteninhalte durch den zentralen ePA-Fachdienst auf Basis des *Cross-Enterprise Document Sharing* (XDS)-Profils der Initiative *Integrating the Healthcare Enterprise* (IHE) verwaltet. Bei der Eröffnung einer ePA wird ein Akten- und ein Kontextschlüssel[3] erzeugt. Diese Schlüssel werden doppelt symmetrisch verschlüsselt für jeden Zugriffsberechtigten in der Komponente Autorisierung eingelagert. Ein hochzuladendes Dokument wird zunächst mit einem zufälligen Dokumentenschlüssel symmetrisch verschlüsselt. Anschließend wird das verschlüsselte Dokument und der wiederum mit dem Aktenschlüssel verschlüsselte Dokumentenschlüssel in der Komponente Dokumentenverwaltung des ePA-Fachdiensts hinterlegt [gem19c].

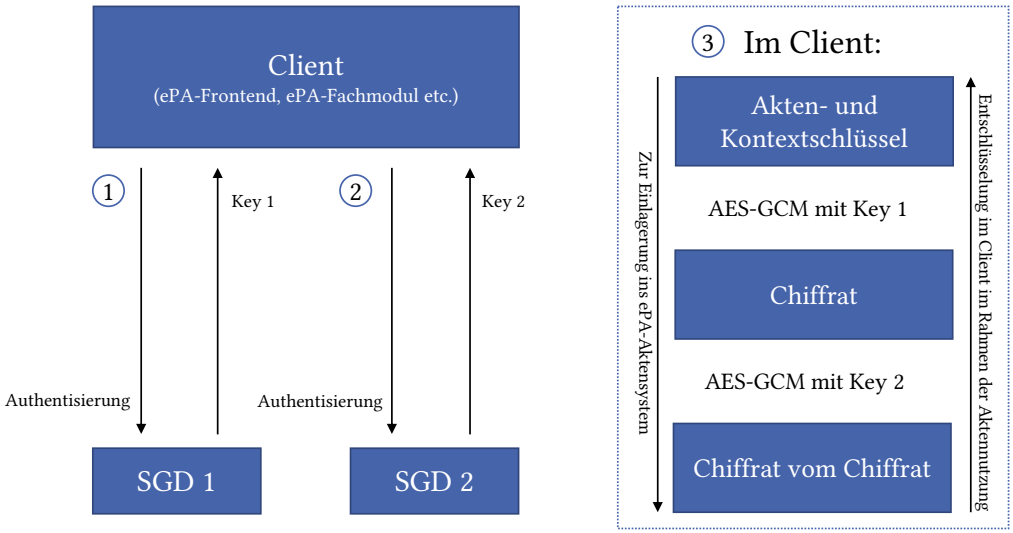

Abbildung 2.2.: Chiffrierung und Dechiffrierung von Schlüsselmaterial im Rahmen der ePA-Nutzung [gem19b]

3 Akten- und Kontextschlüssel sollen regelmäßig ausgetauscht werden, die medizinischen Dokumente werden dabei umgeschlüsselt.

Möchte ein Teilnehmer nun auf ein Dokument zugreifen, muss er sich zunächst mittels Smartcard oder App gegenüber zwei sogenannten Schlüsselgenerierungsdiensten (SGD), SGD 1 und SGD 2, authentisieren, wie in Abbildung 2.2 zu sehen ist. Ein SGD leitet in einem *Hardware Security Module* (HSM) bei erfolgreicher Authentifizierung aus den identifizierenden Merkmalen einen *Advanced Encryption Standard* (AES)-256-Schlüssel ab. Als identifizierende Merkmale werden dabei die Krankenversichertennummer (KVNR) beziehungsweise die gleichbleibende Telematik-ID des Leistungserbringers verwendet, sodass auch nach Verlust einer Smartcard oder über ein Smartphone des Versicherten die Zugriffsmöglichkeit auf die lebenslange ePA sichergestellt ist. Nachdem der Teilnehmer nun von SGD 1 und SGD 2 zwei Schlüssel[4] über einen sicheren Kanal empfangen hat, kann er mit ihnen das von der Komponente Autorisierung („Schlüsselkasten“) erhaltene Chiffrat, welches Akten- und Kontextschlüssel enthält, lokal nach dem Zwiebelschalenprinzip entschlüsseln [gem19b].

Schließlich kann das verschlüsselte Dokument sowie der verschlüsselte Dokumentenschlüssel von der Komponente Dokumentenverwaltung heruntergeladen werden, falls der Teilnehmer laut individuellem Policy-Dokument zugriffsberechtigt[5] ist. Mit dem Aktenschlüssel kann schließlich der Dokumentenschlüssel und damit auch das Dokument entschlüsselt werden. Der Kontextschlüssel erlaubt in der Komponente Dokumentenverwaltung eine serverseitige Suche in den Metadaten, wie Autor, Erstellungszeitpunkt oder Dokumentenformat, unter Berücksichtigung der Zugriffsrechte [gem19c].

2.1.5. Datenspende

Ab der dritten Umsetzungsstufe (1. Januar 2023) sollen Versicherte gemäß § 363 Absatz 1 SGB V-E die Möglichkeit haben, nach einer informierten Einwilligung Inhalte ihrer ePA der medizinischen Forschung zu „spenden“. Den Umfang der Datenfreigabe können Versicherte frei wählen [BK20].

Mit den ePA-Daten soll dabei analog zu den Versorgungsdaten nach §§ 303a ff. SGB V verfahren werden: Der ePA-Provider pseudonymisiert (Pseudonymisierung wird in Abschnitt 3.1 eingeführt) die freigegebenen Daten, versieht sie mit einer Arbeitsnummer und übermittelt sie (verschlüsselt) an das Forschungsdatenzentrum. An die Vertrauensstelle übermittelt der Provider eine Liste, die jeder Arbeitsnummer ein sogenanntes Lieferpseudonym[6] zuordnet. Die Vertrauensstelle leitet anschließend aus den Lieferpseudonymen periodenübergreifende

4 Durch Einsatz verschiedener Master-Keys stimmen die beiden Schlüssel nicht überein.

5 Für das Berechtigungsmanagement wird hierbei der *eXtensible Access Control Markup Language* (XACML)-Standard der *Organization for the Advancement of Structured Information Standards* (OASIS) verwendet.

6 Laut § 303b SGB V ein Versichertenpseudonym, „das eine kassenübergreifende eindeutige Identifizierung im Berichtszeitraum erlaubt“.

Pseudonyme ab und gibt diese mit der zugehörigen Arbeitsnummer an das Forschungsdatenzentrum weiter. Zuletzt verknüpft das Forschungsdatenzentrum die pseudonymisierten Daten über die Arbeitsnummer mit den periodenübergreifenden Pseudonymen. Auch wenn die Forschungsdaten nur einem eingeschränkten Nutzerkreis auf Antrag bereitgestellt werden, scheint die Eignung dieses Verfahrens für die hochsensiblen Akteninhalte wegen der prinzipbedingten Möglichkeit einer Re-Identifizierung von Individuen in pseudonymisierten Datensätzen fragwürdig. Unabhängig davon sollen Versicherte ihre Daten direkt für ein bestimmtes Forschungsvorhaben zur Verfügung stellen können [BT20].

2.2. Datenschutzrechtliche Aspekte

In diesem Abschnitt werden die rechtlichen Grundlagen für den Umgang mit medizinischen Daten behandelt. Bereits im über 2000 Jahren alten Eid des Hippokrates von Kos (Ἱπποκράτης ὁ Κῷος) ist die ärztliche Schweigepflicht formuliert, die heute für Gesundheitsberufe in § 203 Strafgesetzbuch (StGB) verankert und für Ärzte zusätzlich in § 9 Musterberufsordnung der Ärzte (MBO-Ä) standesrechtlich geregelt ist. Fundament des Datenschutzrechts ist jedoch die Datenschutz-Grundverordnung (DSGVO), auf welche nachfolgend eingegangen wird [JS20, S. 23–32].

2.2.1. Europarecht

Die DSGVO wurde zur Harmonisierung der Vorschriften hinsichtlich der Verarbeitung von personenbezogenen Daten verabschiedet und ist seit dem 25. Mai 2018 in allen Mitgliedstaaten der Europäischen Union anzuwenden [JS20, S. 51]. Nach Art. 4 Abs. 1 DSGVO handelt es sich bei personenbezogenen Daten um

> „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden ‚betroffene Person') beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen identifiziert werden kann, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind."

Im Gegensatz zu pseudonymisierten Daten handelt es sich somit bei anonymen Daten nicht um personenbezogene Daten, sie unterliegen also nicht der DSGVO. Gesundheitsdaten sind gemäß

Art. 4 Abs. 15 DSGVO „personenbezogene Daten, die sich auf die körperliche oder geistige Gesundheit einer natürlichen Person, einschließlich der Erbringung von Gesundheitsdienstleistungen, beziehen und aus denen Informationen über deren Gesundheitszustand hervorgehen“ [JS20, S. 54–64; DR20, S. 164–168].

Da Gesundheitsdaten zu den besonderen Kategorien personenbezogener Daten gehören, dürfen sie laut Art. 9 Abs. 1 DSGVO mit Ausnahme der in Abs. 2 genannten Fälle prinzipiell nicht verarbeitet werden (sog. Verbot mit Erlaubnisvorbehalt). So ist eine Verarbeitung beispielsweise erlaubt, wenn eine explizite Einwilligung des Betroffenen vorliegt, sie im Rahmen der Behandlung durch Berufsgeheimnisträger oder für Forschungszwecke gemäß Art. 89 DSGVO erfolgt. Hierbei stellt Art. 89 Abs. 1 DSGVO recht umfangreiche Forderungen, weshalb man in aller Regel mit der Ausnahme des Verarbeitungsverbots bei informierter Einwilligung arbeitet:[7]

> “Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.”

Die Daten sind also vorrangig zu anonymisieren. Nur wenn auf diese Weise die Forschungszwecke nicht erreicht werden können, ist eine Pseudonymisierung ausreichend [Küh19; DR20, S. 168–173].

2.2.2. Bundesrecht

Historisch bedingt ist der Datenschutz in Deutschland von großer Bedeutung. Dieser Umstand wird bereits durch das Grundrecht auf informationelle Selbstbestimmung deutlich, zu welchem das Volkszählungsurteil beigetragen hat [PS17, S. 140–144]. Als europäische Verordnung hat die DSGVO jedoch Vorrang gegenüber nationalem Recht. Durch sogenannte Öffnungsklauseln können die Mitgliedstaaten Ergänzungen und Konkretisierungen vornehmen, welche in Deutschland im Bundesdatenschutzgesetz (BDSG) festgesetzt sind.[8] Das BDSG wird allerdings

7 Die englische Fassung ist an dieser Stelle prägnanter, siehe hierzu S. 169 in [DR20].

8 Eine solche Öffnungsklausel gibt es beispielsweise in Art. 89 DSGVO, die von § 27 BDSG ausgefüllt wird. Es sei auf die Einschätzung von Dierks und Roßnagel in [DR20, S. 164–173] hingewiesen, ob die in § 27 BDSG geregelten Vorgaben zur Anonymisierung rechtswirksam sind.

durch bereichsspezifische Regelungen verdrängt („*lex specialis derogat legi generali*“) [JS20, S. 52–54].

Zu den bereichsspezifischen Datenschutznormen im Gesundheitswesen gehören auf Bundesebene beispielsweise das Sozialgesetzbuch (SGB) und das Infektionsschutzgesetz (IfSG) (siehe insbesondere § 14 IfSG). Auf Ebene der Länder enthalten zum Beispiel die Krankenhaus- oder Krebsregistergesetze datenschutzrechtliche Vorschriften [Küh19].

2.2.3. Standard-Datenschutzmodell

Das Standard-Datenschutzmodell (SDM) ist nach der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder ein Werkzeug, „mit dem die Auswahl und Bewertung technischer und organisatorischer Maßnahmen unterstützt wird, die sicherstellen und den Nachweis dafür erbringen, dass die Verarbeitung personenbezogener Daten nach den Vorgaben der DS-GVO erfolgt“ [DSK20]. Es basiert auf sieben Schutzzielen, welche nachfolgend vorgestellt werden:

Nichtverkettung Dieser Grundsatz besagt, dass personenbezogene Daten nicht zusammengeführt werden dürfen. Eine Datenverarbeitung ist also nur für einen festen Zweck erlaubt [BEE18, S. 39 f.].

Datenminimierung Nach dem Schutzziel der Datenminimierung muss die Datenverarbeitung auf das für den jeweiligen Zweck unbedingt notwendige Maß beschränkt werden [DSK20].

Verfügbarkeit Verfügbarkeit ist gegeben, wenn autorisierte Nutzer jederzeit auf die Daten zugreifen können. Aus diesem Schutzziel ergibt sich die Notwendigkeit einer regelmäßigen Datensicherung [BEE18, S. 40 f.].

Integrität Unter Integrität wird der Schutz vor unautorisierter Veränderung der zu verarbeitenden Daten beispielsweise mithilfe digitaler Signaturen oder des Mehr-Augen-Prinzips verstanden [DSK20].

Vertraulichkeit Dieses Schutzziel ist gewährleistet, wenn keine unbefugte Nutzung der Datensätze möglich ist. In diesem Zusammenhang ist ein auf den Anwendungsfall zugeschnittenes Berechtigungsmanagement entscheidend [PS17, S. 10].

Transparenz Damit das Transparenzgebot erfüllt ist, müssen betroffene Personen und Kontrollinstanzen die Datenverarbeitung nachvollziehen können. Nur auf diese Weise kann eine informierte Einwilligung erfolgen [BEE18, S. 36 f.].

Intervenierbarkeit Intervenierbarkeit ist erfüllt, wenn Betroffene die ihnen zustehenden Rechte ausüben können. Zu diesen sogenannten Betroffenenrechten gehört beispielsweise das Auskunftsrecht, das Recht auf Löschung und das Recht auf Datenübertragbarkeit [DSK20].

2.3. Verwandte Arbeiten

Für die im nächsten Kapitel dargelegten *Privacy-Enhancing Technologies* (PETs) finden sich in der Literatur bereits unzählige Algorithmen und raffinierte Implementierungen. Dem Autor ist jedoch keine (nicht kommerzielle) Arbeit bekannt, die verschiedene PETs im Rahmen einer modernen Forschungsschnitstelle für medizinische Daten nutzbar macht.

Lablans, Borg und Ückert [LBÜ15] stellen in ihrer Publikation eine Pseudonymisierungsschnittstelle namens Mainzelliste[9] vor, welche identifizierenden Merkmalen Pseudonyme zuordnet. Eine feste Kombination identifizierender Daten wird dabei auf ein gleichbleibendes Pseudonym abgebildet, sodass Einträge aus verschiedenen Datenquellen, die einer bestimmten Person zuzuordnen sind, zusammengeführt werden können. Die Schnittstelle ist wegen des REST-basierten Ansatzes für die Verwendung in Webanwendungen prädestiniert.

Im Bereich der Anonymisierung wird das performante Tool *ARX*[10] in Forschungsprojekten verschiedener Universitäten eingesetzt. Wie aus einem Artikel von Prasser et al. [Pra+20] hervorgeht, bietet *ARX* ein wesentlich breiteres Spektrum an Anonymisierungstechniken als vergleichbare Open-Source-Lösungen. Gleichzeitig übertrifft es verwandte Tools in puncto Skalierbarkeit und Qualität der Ausgabedaten.

Hinsichtlich der *Differential Privacy* (DP) sind die Bibliotheken[11] von Google zu nennen. In einem Artikel von Wilson et al. [Wil+20] wird der technische Hintergrund der PostgreSQL-Erweiterung für DP erläutert. Insbesondere wird darin beleuchtet, wie man dem Problem begegnen kann, dass einem Nutzer mehrere Datenbankeinträge zugeordnet werden können. Mit *OpenDP*[12] hat auch die Harvard University in Zusammenarbeit mit Microsoft vor Kurzem ein Toolkit für DP veröffentlicht.

Die *Microsoft Simple Encrypted Arithmetic Library*[13] für homomorphe Verschlüsselung ermöglicht Berechnungen auf verschlüsselten Daten. Gilad-Bachrach et al. [Gil+16] beschreiben in ihrem Konferenzbeitrag, wie mit dieser Bibliothek verschlüsselte Daten unter Verwendung von neuronalen Netzen ausgewertet werden können. Am Beispiel des *Modified National Institute of Standards and Technology* (MNIST)-Datensatzes zeigen die Autoren, dass mit einem bereits trainierten neuronalen Netz eine Genauigkeit von 99 Prozent bei der Erkennung von handgeschriebenen Ziffern in verschlüsselter Form erreicht werden konnte. Im Bereich der medizinischen Forschung wurden bereits verschiedene homomorphe Verschlüsselungsverfahren evaluiert, nachfolgend werden drei Arbeiten kurz vorgestellt: Dowlin et al. [Dow+17]

9 `https://bitbucket.org/medicalinformatics/mainzelliste` (besucht am 22. 07. 2020)
10 `https://github.com/arx-deidentifier/arx` (besucht am 07. 07. 2020)
11 `https://github.com/google/differential-privacy` (besucht am 08. 07. 2020)
12 `https://github.com/opendifferentialprivacy` (besucht am 07. 07. 2020)
13 `https://github.com/microsoft/SEAL` (besucht am 09. 07. 2020)

legen in ihrer Publikation dar, wie biomedizinische Daten mit der zuvor genannten Bibliothek analysiert werden können. Kocabas et al. [Koc+13] zeigen, dass einfache Auswertungen wie die Berechnung der durchschnittlichen Herzfrequenz in Echtzeit realisierbar sind, auch wenn sie einen stark erhöhten Speicherbedarf beobachten. Vizitiu et al. [Viz+19] verwenden ein unter dem Namen *Matrix Operation for Randomization and Encryption* bekanntes *Fully Homomorphic Encryption Scheme*, um Aufnahmen aus der Koronarangiografie mit einem neuronalen Netz zu klassifizieren. Ihre Ergebnisse weisen im Vergleich zur unverschlüsselten Version eine identische Klassifikationsgenauigkeit bei einer um Faktor 33 verlängerten Laufzeit auf.

Derzeit befindet sich im Rahmen des EU-Forschungsprogramms *Horizon 2020* das *Platform for Privacy Preserving Data Analytics* (PAPAYA)-Framework[14] in der Entwicklung, welches datenschutzgerechte Datenanalysen erlauben soll. Konkret soll PAPAYA die Möglichkeit bieten, Training und Klassifizierung von neuronalen Netzen, Clusteranalysen sowie grundlegende statistische Auswertungen DSGVO-konform durchzuführen. Hierbei soll *Secure Multiparty Computation*, DP und homomorphe Verschlüsselung Anwendung finden. PAPAYA soll unter anderem anhand zweier medizinischer Anwendungsfälle, der Erkennung von Herzrhythmusstörungen sowie von Stress, validiert werden. Im Unterschied zu PAPAYA liegt der Schwerpunkt dieser Arbeit nicht auf maschinellem Lernen [PAP19; Cic+19].

14 `https://www.papaya-project.eu` (besucht am 08. 07. 2020)

3. PETs im Gesundheitswesen

Technische Methoden zum Schutz der Privatsphäre werden in der Literatur gemeinhin als *Privacy-Enhancing Technologies* (PETs) bezeichnet. In diesem Kapitel werden die wichtigsten PETs vorgestellt. Das Hauptaugenmerk liegt dabei auf den verschiedenen Anonymisierungstechniken, die den Zielkonflikt zwischen Datenschutz und Forschungsdatennutzung bestmöglich zu lösen versuchen. Gleichwohl bleiben die sich durch Pseudonymisierung und homomorphe Verschlüsselungsverfahren ergebenden Möglichkeiten nicht unerwähnt. Dabei lassen sich Pseudonymisierung und Anonymisierung primär auf Datensätze anwenden, die in einer Datenbank abgespeichert sind, während der Einsatz homomorpher Verschlüsselungsverfahren prinzipiell bei allen Dateiformaten möglich ist.

3.1. Pseudonymisierung

Unter dem Begriff der Pseudonymisierung versteht man eine Prozedur, bei der durch eine Zuordnungsvorschrift personenbezogene Daten so abgeändert werden, dass ein Personenbezug ohne ebenjene Zuordnungsvorschrift erschwert oder ausgeschlossen ist. Identifizierende Merkmale wie Name oder Adresse (siehe hierzu auch Abschnitt 3.2.2) werden also durch ein beliebiges, aber festes Kennzeichen, das Pseudonym, ersetzt. Die Zuordnungsvorschrift wird in der Regel so gewählt, dass eine Re-Identifizierung prinzipiell möglich ist. Beispielsweise können Personen so bei Bedarf kontaktiert werden, um zusätzliche Daten zu erheben oder über Studienergebnisse zu informieren. Insgesamt lassen sich drei verschiedene Arten von Pseudonymen unterscheiden [PS17; PR04]:

Selbstgenerierte Pseudonyme Hierbei wählt der Betroffene das Pseudonym selbst. Folglich kann auch nur der Betroffene den Personenbezug wiederherstellen [BLD97].

Referenzpseudonyme Den Identitätsdaten werden mithilfe einer sogenannten Referenzliste Pseudonyme zugeordnet, indem man die Pseudonyme einfach durchnummeriert oder pseudozufällig generiert. Diese Aufgabe wird oftmals von einer *Trusted Third Party* (TTP) übernommen. Eine Wiederherstellung des Personenbezugs ist nur über die Referenzliste möglich, welche nach der DSGVO getrennt von den pseudonymisierten Daten aufzubewahren ist [BLD97; PR04; BEE18, S. 157].

Einwegpseudonyme Durch parametrisierbare Einwegfunktionen lassen sich Identitätsdaten in Pseudonyme überführen. Im Unterschied zu Referenzpseudonymen basiert die Sicherheit nicht auf der Geheimhaltung der Abbildungsvorschrift, sondern auf der Geheimhaltung der Funktionsparameter (Kerckhoffs'sches Prinzip). Für eine Re-Identifizierung müssen sowohl die Funktionsparameter als auch die Identitätsdaten bekannt sein. Als Einwegfunktion kommt beispielsweise ein *Keyed-Hash Message Authentication Code* (HMAC) infrage, der die Erzeugung eines Hashwerts von einem zusätzlichen geheimen Schlüssel abhängig macht [BLD97; SW19].

Ferner können auch mehrere unabhängige Stellen einen Datensatz hintereinander pseudonymisieren, um die Sicherheit eines Pseudonymisierungsverfahrens zu erhöhen. In Deutschland findet die Pseudonymisierung unter anderem bei Krebsregistern Anwendung, um verschiedene Behandlungsmethoden vergleichen zu können [SW19].

3.2. Anonymisierung

Einen ersten Hinweis für die Bedeutung des Begriffs Anonymität liefert die (alt-)griechische Sprache: Anonymos (ανώνυμος) kann mit „namenlos, ungenannt" übersetzt werden. Wie bereits erwähnt, sind nach der DSGVO anonyme Daten das Gegenteil von personenbezogenen Daten. Im Unterschied zur Pseudonymisierung soll also durch Anonymisierung, worunter man Verfahren zur Herstellung von Anonymität versteht, eine Wiederherstellung des Personenbezugs praktisch ausgeschlossen werden können. Genauer soll eine Re-Identifizierung nur mit unverhältnismäßig großem Aufwand möglich sein. Eine absolute Anonymisierung, die eine Re-Identifizierung von Individuen zweifelsfrei ausschließt, ist für gewöhnlich nicht möglich und nach Ansicht des Bundesbeauftragten für den Datenschutz und die Informationsfreiheit datenschutzrechtlich auch nicht geboten. Nachfolgend werden zunächst Szenarien betrachtet, in welchen Anonymisierung Verwendung findet. Anschließend werden die verschiedenen Arten von Attributen, aus denen sich ein personenbezogenes Datum zusammensetzt, vorgestellt und Bedrohungen erläutert, die mit der Bereitstellung eines Datensatzes zu Forschungszwecken einhergehen. Abschließend werden die wichtigsten Anonymitätskriterien und Anonymisierungsverfahren behandelt [PS17, S. 12 f.; BfDI20].

3.2.1. Klassifikation

Grundsätzlich sind das interaktive und das nichtinteraktive Szenario der Forschungsdatennutzung voneinander zu trennen. Bei letzterem wird nach entsprechender Anonymisierung der gesamte Datensatz weitergegeben, weshalb dieses Setting auch als *Privacy-Preserving Data Publishing* (PPDP) bekannt ist. Der Empfänger kann also zeitlich unbefristet beliebige

Analysen auf den Daten durchführen. Daher muss auch prinzipiell damit gerechnet werden, dass sich Unbefugte Zugriff auf den Datensatz verschaffen. Gängige Anonymitätsmaße für nichtinteraktive Verfahren sind k-Anonymity, ℓ-Diversity, t-Closeness und δ-Presence [PS17, S. 28; Fun+10].

Das interaktive Szenario beschreibt dagegen eine Datennutzung, bei der externe Nutzer über eine Schnittstelle nur bestimmte Anfragen stellen können. Dieses Setting ist in der Literatur auch unter den Termini *Privacy-Preserving Data Mining* (PPDM) oder *Statistical Database* geläufig. Im Vergleich zum zuvor dargelegten Szenario besteht für die Nutzer folglich keine direkte Zugriffsmöglichkeit auf die Daten, sie erhalten lediglich die – möglicherweise verrauschten – Ergebnisse zu ihren Anfragen. Dass die Einrichtung, welche die Schnittstelle anbietet, die Datenhoheit behält, ist zudem ein wesentlicher Vorteil dieses Ansatzes. Vorteilhaft ist des Weiteren, dass die zum Einsatz kommende Implementierung auf Basis einer PET jederzeit ausgetauscht werden kann, beispielsweise um auf einen neuen Angriffsvektor zu reagieren. Allerdings sind die Analysemethoden in diesem Szenario häufig stark eingeschränkt, meist können nur Aggregatfunktionen in den Anfragen verwendet werden. Hierzu gehören unter anderem *Counting Queries*, welche die Anzahl ein bestimmtes Kriterium erfüllender Individuen berechnen, oder *Predicate Queries*, die nichts anderes als auf den Wertebereich zwischen Null und Eins normierte *Counting Queries* sind. Das wichtigste Konzept für eine interaktive Datennutzung ist DP [DE13; GLS14; Zig+20].

3.2.2. Attributtypen und Notation

Personenbezogene medizinische Daten sind in elektronischen Datenverarbeitungssystemen normalerweise in relationalen Datenbanken organisiert. Eine solche Datenbank kann man sich in der Notation nach Machanavajjhala et al. [Mac+07] wie eine Tabelle T vorstellen, die wiederum aus m Spalten, den sogenannten Attributen $A_1, A_2, \ldots, A_m$, und n Zeilen, den

Name	KVNR	PLZ	Geburtsdatum	Geschlecht	Diagnose
Xaver Mayers	C753573156	76139	08.10.1981	männlich	Heuschnupfen
Jule Steiner	A676007843	76133	30.03.1981	weiblich	Asthma
Wilhelm Dürer	E392080529	76133	14.12.1989	männlich	Diabetes
Simone Müller	G967985715	76149	26.06.1981	weiblich	Erkältung
Jutta Weinert	H128962713	76133	02.08.1981	weiblich	Borreliose
Dietmar Hees	F432681510	76131	19.01.1985	männlich	Tinnitus
Philipp Fischer	B516329096	76133	27.04.1983	männlich	Heuschnupfen
Nicole Pfeifer	D228851627	76135	11.12.1981	weiblich	Erkältung

Tabelle 3.1.: Verschiende Typen von Attributen in einer Datenbank

sogenannten Tupeln $t_1, t_2, \ldots, t_n$, besteht. Die Menge aller Attribute $\{A_1, A_2, \ldots, A_m\}$ sei mit $\mathcal{A}$ bezeichnet, für $C = \{C_1, C_2, \ldots, C_p\} \subseteq \mathcal{A}$ sei $t[C]$ die Projektion eines Tupels t auf die Attribute in C.

Tabelle 3.1 zeigt einen Auszug einer beispielhaften Datenbank, wobei sich jedes Tupel $t_i \in T$ einem Patienten zuordnen lässt. Bei den Attributen sind nun verschiedene Typen zu unterscheiden:

Direkte Identifikatoren Sie erlauben eine unmittelbare Re-Identifizierung von Personen, im Beispiel `Name` und `KVNR`.

Quasi-Identifikatoren Ein Quasi-Identifikator ist demgegenüber eine Menge von Attributen, die in Verbindung mit externen Informationen, zum Beispiel aus einem Telefonbuch oder den sozialen Medien, zu einer Re-Identifizierung führt. Im Kontext der Anonymisierung sollten alle Attribute als quasi-identifizierend angesehen werden, die ein Angreifer – möglicherweise gesetzeswidrig – in externen Quellen einsehen könnte [Fun+10]. Welche Gefahr Quasi-Identifikatoren beispielsweise für Patientendaten darstellen, zeigte Latanya Sweeney [Swe02] in bemerkenswerter Weise: Ihr gelang es, in einem vorgeblich anonymisierten Datensatz der *Massachusetts Group Insurance Commission*, die für die Krankenversicherung von 135 000 Staatsbediensteten und deren Familien zuständig ist, durch eine Verknüpfung mit einem öffentlichen Wählerverzeichnis den damaligen Gouverneur von Massachusetts, William Weld, eindeutig zu identifizieren. Die Schnittmenge der Attribute von beiden Datensätzen und damit Quasi-Identifikator war hierbei die Kombination aus `Postleitzahl`, `Geburtsdatum` und `Geschlecht`, wie Abbildung 3.1 veranschaulicht.

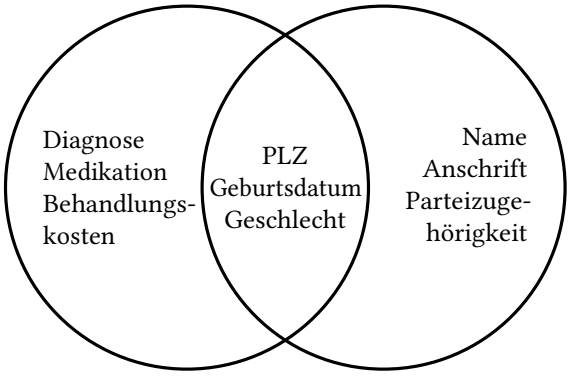

Abbildung 3.1.: Verknüpfung zweier Datensätze über gemeinsame Attribute [Swe02]

Sensible Attribute Derartige Attribute enthalten schließlich besonders schützenswerte Informationen, die bei unbefugter Offenlegung einer Person unmittelbar schaden könnten. In Tabelle 3.1 ist das Attribut `Diagnose` ein sensibles Attribut [GLS14].

3.2.3. Bedrohungen

In der Literatur werden drei Arten von Bedrohungen für die Privatsphäre von Individuen beschrieben, die bei der Veröffentlichung eines Datensatzes berücksichtigt werden müssen:

Identity Disclosure (auch Re-Identifizierung) Bei der Re-Identifizierung gelingt es einem Angreifer, einen bestimmten Eintrag des Datensatzes einer Person zuzuordnen [GLS14]. Sweeney hat einige solcher Offenlegungen durchgeführt, um ein Bewusstsein für derartige Angriffe zu schaffen (siehe Beispiel aus letztem Abschnitt). Weithin bekannt ist auch der Konferenzbeitrag von Narayanan und Shmatikov [NS08] zum *Netflix Prize Dataset*, welchen Netflix im Rahmen eines Wettbewerbs zur Verbesserung seines Empfehlungsalgorithmus veröffentlicht hatte. Den Autoren gelang es, auf Netflix abgegebene Filmrezensionen mit entsprechenden Bewertungen in der *Internet Movie Database*, die oftmals unter Klarnamen verfasst wurden, zu verknüpfen und so den von Netflix bei der Veröffentlichung entfernten Personenbezug wiederherzustellen.

Attribute Disclosure Auch ohne direkte Zuordnung eines Eintrages zu einem Individuum kann ein Angreifer möglicherweise das Individuum mit dem Wert eines sensiblen Attributs assoziieren. Dieser Fall kann beispielsweise eintreten, wenn alle Einträge, die für eine Person infrage kommen, denselben Wert des sensiblen Attributs aufweisen [GL15, S. 115].

Membership Disclosure Allein die Tatsache, dass ein Individuum mit hoher Wahrscheinlichkeit im Datensatz enthalten ist (oder eben nicht), stellt eine ernsthafte Bedrohung dar. Im Beispiel einer unzureichend geschützten Datenbank eines Krebsregisters könnte ein Angreifer ermitteln, ob eine bestimme Person Krebs hat oder nicht [GLS14].

3.2.4. Anonymitätsmaße

Es gibt zahlreiche Anonymitätskriterien, um die soeben vorgestellten Bedrohungen abzuwehren. Im Zuge einer Anonymisierung müssen zunächst alle direkten Identifikatoren entfernt werden, andernfalls muss definitionsgemäß mit einer Re-Identifizierung einzelner Individuen gerechnet werden. Tabelle 3.2 zeigt das Ergebnis dieses Vorgangs für die Patientendaten aus Tabelle 3.1 [Fun+10].

In den nachfolgenden Abschnitten werden ausgewählte Anonymitätsmaße für das nichtinteraktive Szenario vorgestellt, anschließend *Differential Privacy* (DP) als bedeutendstes Kriterium der interaktiven Forschungsdatennutzung. Für eine detaillierte Auseinandersetzung mit diesem Themengebiet sei auch auf die Übersichtsarbeiten von Fung et al. [Fun+10] und Zigomitros et al. [Zig+20] verwiesen.

PLZ	Geburtsdatum	Geschlecht	Diagnose
76139	08.10.1981	männlich	Heuschnupfen
76133	30.03.1981	weiblich	Asthma
76133	14.12.1989	männlich	Diabetes
76149	26.06.1981	weiblich	Erkältung
76133	02.08.1981	weiblich	Borreliose
76131	19.01.1985	männlich	Tinnitus
76133	27.04.1983	männlich	Heuschnupfen
76135	11.12.1981	weiblich	Erkältung

Tabelle 3.2.: Patientendaten nach Entfernen der direkten Identifikatoren

3.2.4.1. k-Anonymity

Samarati und Sweeney präsentierten in [SS98] erstmals ein Konzept, um die Re-Identifizierung einzelner Individuen durch einen herausgegebenen Datensatz zu vermeiden. In Abhängigkeit des Parameters k gewährleistet k-Anonymity, dass eine Person hinsichtlich des Quasi-Identifikators von mindestens $k-1$ anderen Personen ununterscheidbar ist. Je höher der Wert von k ist, desto geringer ist also die Wahrscheinlichkeit einer Re-Identifizierung. Die Menge aller Tupel, die in einer Wertekombination der quasi-identifizierenden Attribute übereinstimmen, bildet hierbei eine sogenannte Äquivalenzklasse. Formal ergibt sich die folgende Definition [Ema+09]:

Definition 3.1 (k-Anonymity) Sei QI_T die Menge aller quasi-identifizierenden Attribute einer Tabelle T. Dann ist T k-anonym, wenn für jedes Tupel $t \in T$ mindestens $k-1$ weitere Tupel $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$ existieren, sodass gilt [Swe02; Mac+07]:

$$t[QI_T] = t_{i_1}[QI_T] = t_{i_2}[QI_T] = \cdots = t_{i_{k-1}}[QI_T].$$

Durch Generalisierung (siehe Abschnitt 3.2.5.1) erhält man mit Tabelle 3.3 eine k-anonyme Version der Patientendaten aus Tabelle 3.2, wobei $k = 4$ und $QI_T = \{$`PLZ`, `Geburtsdatum`, `Geschlecht`$\}$. Die anonymisierte Tabelle weist dabei zwei Äquivalenzklassen auf, deren jeweilige Vertreter – repräsentiert durch die ersten und letzten vier Einträge – in den quasi-identifizierenden Attributen übereinstimmen. Selbst wenn ein Angreifer `PLZ`, `Geburtsdatum` und `Geschlecht` eines Individuums kennt, kommen immer noch alle Vertreter der Äquivalenzklasse, also im konkreten Fall vier Einträge, infrage.

Wird ein Datensatz unter Verwendung von k-Anonymity wiederholt anonymisiert und veröffentlicht, müssen mit *Unsorted Matching* und *Complementary Release* zwei Bedrohungen berücksichtigt werden, die das Anonymitätsmaß aushebeln können. Die erstgenannte Bedrohung ergibt sich aus der Tatsache, dass die Einträge einer realen Datenbank fast immer

PLZ	Geburtsdatum	Geschlecht	Diagnose
7613∗	∗∗.∗∗.198∗	männlich	Heuschnupfen
7613∗	∗∗.∗∗.198∗	männlich	Diabetes
7613∗	∗∗.∗∗.198∗	männlich	Tinnitus
7613∗	∗∗.∗∗.198∗	männlich	Heuschnupfen
761∗∗	∗∗.∗∗.1981	weiblich	Borreliose
761∗∗	∗∗.∗∗.1981	weiblich	Erkältung
761∗∗	∗∗.∗∗.1981	weiblich	Asthma
761∗∗	∗∗.∗∗.1981	weiblich	Erkältung

Tabelle 3.3.: Beispiel einer 4-anonymen Tabelle

in einer bestimmten Reihenfolge vorliegen, auch wenn das relationale Datenmodell an sich keine Reihenfolge bei den Tupeln kennt. Auf diese Weise können unbeabsichtigt Informationen publik werden, falls eine Tabelle mehrfach in k-anonymer Form veröffentlicht wird. So stellen die mittlere und die rechte Tabelle in Abbildung 3.2 2-anonyme Versionen der linken Tabelle dar, die durch Verknüpfung wieder die linke, ursprüngliche Tabelle ergeben – unter der Voraussetzung, dass die Tupel in beiden Tabellen in der gleichen Reihenfolge stehen. Durch Umsortieren der Tupel vor der Veröffentlichung einer Tabelle lässt sich ein derartiger Angriff verhindern [Swe02].

PLZ	Geschlecht
76149	männlich
76149	weiblich
76133	weiblich
76133	männlich

PLZ	Geschlecht
761∗∗	männlich
761∗∗	weiblich
761∗∗	weiblich
761∗∗	männlich

PLZ	Geschlecht
76149	∗
76149	∗
76133	∗
76133	∗

Abbildung 3.2.: Bedrohung durch *Unsorted Matching* [Swe02]

Bilden veröffentlichte Versionen einer Tabelle ein *Complementary Release*, so ist ein alleiniges Umsortieren nicht hinreichend, um Individuen durch k-Anonymity vor einer Re-Identifizierung zu schützen. Ursächlich für den genannten Schwachpunkt ist, dass Tupel auch über Attribute, die zunächst nicht als quasi-identifizierend wahrgenommen werden, mit Tupeln einer anderen Tabelle verknüpft werden können. Beispielsweise lässt sich Tabelle 3.4, eine ebenfalls 4-anonyme Variante der ursprünglichen Tabelle 3.2, über das sensible Attribut `Diagnose` mit Tabelle 3.3 verbinden. Da nicht alle Personen mit der Postleitzahl `76133` in der resultierenden Tabelle das gleiche Geschlecht haben,[1] ist 4-Anonymity nun verletzt. Diese Verwundbarkeit kann vermieden werden, indem alle bereits herausgegebenen Attribute bei nachfolgenden

1 Die Diagnose `Diabetes` lässt sich einer männlichen Person zuordnen, `Borreliose` hingegen einer weiblichen.

Veröffentlichungen in die Menge der quasi-identifizierenden Attribute QI_T aufgenommen werden. Infolgedessen wäre Tabelle 3.4 auch keine valide 4-Anonymisierung [PS17, S. 35 f.].

PLZ	Geburtsdatum	Geschlecht	Diagnose
76133	**.**.198*	*	Heuschnupfen
76133	**.**.198*	*	Diabetes
76133	**.**.198*	*	Borreliose
76133	**.**.198*	*	Asthma
761**	**.**.198*	*	Tinnitus
761**	**.**.198*	*	Heuschnupfen
761**	**.**.198*	*	Erkältung
761**	**.**.198*	*	Erkältung

Tabelle 3.4.: Bedrohung durch *Complementary Release*

3.2.4.2. ℓ-Diversity

Auch bei richtiger Anwendung schützt k-Anonymity nicht vor *Attribute Disclosure*, wie zwei Angriffe, *Homogeneity Attack* und *Background Knowledge Attack*, zeigen. Ersterer Angriff ist möglich, wenn alle Vertreter einer Äquivalenzklasse bei einem sensiblen Attribut den gleichen Wert besitzen, also homogen sind. Falls ein Angreifer weiß, in welcher Äquivalenzklasse sich eine Person befindet, so kann er auf diese Weise den Wert des sensiblen Attributs herausfinden. Die 3-anonyme Tabelle 3.5 illustriert diesen Sachverhalt: Ist dem Angreifer bekannt, dass die Zielperson im Jahr `1932` geboren wurde, `männlich` und in `76228` wohnhaft ist, kann er daraus auf die Diagnose `Krebs` schließen [Mac+07].

PLZ	Geburtsjahr	Geschlecht	Diagnose
...	...	...	...
7622*	1932	weiblich	Erkältung
7622*	1932	männlich	Krebs
7622*	1932	männlich	Krebs
7622*	1932	männlich	Krebs

Tabelle 3.5.: *Homogeneity Attack*

Wie bereits aus dem Namen *Background Knowledge Attack* hervorgeht, setzt die zweite Angriffsmöglichkeit Hintergrundwissen voraus. Ein Angreifer kann unter Anwendung dieses Wissens jedem Wert eines sensiblen Attributs – sofern er die Äquivalenzklasse eines Individuums kennt – eine gewisse Wahrscheinlichkeit zuordnen und sodann unwahrscheinliche

Vertreter ausschließen. Liegt einem Angreifer zum Beispiel Tabelle 3.6 vor, kann er für die Zielperson, welche im Jahr `1954` geboren wurde, `männlich` und in `76229` wohnhaft ist, mit dem Hintergrundwissen, dass `Brustkrebs` bei Männern äußerst selten auftritt, nach dem Ausschlussprinzip die Diagnose `Diabetes` folgern [LLV07; PS17, S. 37].

PLZ	Geburtsjahr	Geschlecht	Diagnose
...	...	...	...
7622∗	1953	∗	Erkältung
7622∗	1954	∗	Brustkrebs
7622∗	1954	∗	Diabetes
7622∗	1954	∗	Brustkrebs

Tabelle 3.6.: *Background Knowledge Attack*

Mit ℓ-Diversity stellten Machanavajjhala et al. [Mac+07] ein Anonymitätsmaß vor, das vor beiden Angriffen schützt. Leitgedanke ist dabei die Gewährleistung ℓ verschiedener Ausprägungen eines sensiblen Attributs innerhalb der Äquivalenzklassen [Fun+10].

Definition 3.2 (Prinzip der ℓ-Diversity) Eine Äquivalenzklasse ist ℓ-divers, wenn die Vertreter gemäß einer konkreten ℓ-Diversity-Instanz mindestens ℓ „gut repräsentierte“ Werte für das sensible Attribut aufweisen. Eine Tabelle ist ℓ-divers, wenn alle Äquivalenzklassen ℓ-divers sind [LLV07].

Offensichtlich begegnet man mit dieser Definition dem Problem der Homogenität, da in jeder Äquivalenzklasse unterschiedliche Attributwerte vorliegen. Außerdem muss ein Angreifer nun $\ell - 1$ „gut repräsentierte“ Werte des sensiblen Attributs durch Hintergrundwissen ausschließen können, um den tatsächlichen Wert aufzudecken. Ein ℓ-diverser Datensatz ist darüber hinaus immer auch k-anonym, wobei $k = \ell$. Für den Ausdruck „gut repräsentiert“ gibt es in diesem Zusammenhang mehrere Instanzen, im Folgenden werden lediglich drei ausgewählte Alternativen erläutert [Mac+07]:

Distinct ℓ-Diversity Eine Tabelle entspricht diesem Kriterium, wenn jede Äquivalenzklasse in Bezug auf das sensible Attribut mindestens ℓ verschiedene Werte enthält. Nach dieser Definition ist zum Beispiel Tabelle 3.4 3-divers. Problematisch an diesem Ansatz ist, dass ein Angreifer weiterhin signifikante Information gewinnen kann, wenn bestimmte Attributwerte innerhalb einer Äquivalenzklasse besonders häufig auftreten [GLS14].

Entropy ℓ-Diversity Sei die Domäne des sensiblen Attributs, also die Menge aller möglichen Werte, mit S bezeichnet und $p_{(E,s)}$ sei der Anteil der Tupel einer Äquivalenzklasse E mit dem

Wert s. Dann lässt sich E im Sinne der Informationstheorie die Entropie

$$H(E) = -\sum_{s \in S} p_{(E,s)} \log p_{(E,s)}$$

zuordnen. Ein Datensatz ist vor diesem Hintergrund ℓ-divers, wenn für jede Äquivalenzklasse E gilt: $H(E) \geq \log \ell$. Damit fordert diese Definition nicht nur mindestens ℓ verschiedene Werte für das Attribut innerhalb jeder Äquivalenzklasse, sondern auch eine gleichmäßige Verteilung[2] der Attributwerte [LLV07].

Recursive (c, ℓ)-Diversity Diese Variante von ℓ-Diversity stellt sicher, dass besonders häufige Werte des sensiblen Attributs nicht zu häufig auftreten und besonders seltene nicht zu selten. Sei m die Anzahl der Ausprägungen dieses Attributs in der Äquivalenzklasse, ferner sei r_i die absolute Häufigkeit des i-ten Werts bei absteigender Sortierung nach Häufigkeit. Dann ist eine Tabelle rekursiv (c, ℓ)-divers, wenn gilt: $f_1 < c \sum_{i=l}^{m} f_i$. Der häufigste Wert muss also seltener sein als die mit einer festgelegten Konstante c multiplizierte Summe der $m - l + 1$ seltensten Werte [Fun+10].

3.2.4.3. t-Closeness

ℓ-Diversity weist zwei Schwächen auf, weshalb kein vollumfänglicher Schutz vor *Attribute Disclosure* besteht: Voraussetzung für den ersten Angriff, *Skewness Attack*, ist eine schiefe Verteilung der sensiblen Attributwerte. Ein Angreifer habe beispielsweise Zugriff auf eine Tabelle, welche die Ergebnisse der Tests auf ein beliebiges Virus enthält, außerdem sei ihm die Äquivalenzklasse der Zielperson geläufig. Hierbei seien 99 Prozent der Ergebnisse in der gesamten Tabelle negativ. Falls die Äquivalenzklasse der Zielperson gleich viele positive wie negative Einträge besitzt, erfüllt sie zwar alle zuvor genannten Instanzen von 2-Diversity. Jedoch hat dieser Umstand den unerwünschten Nebeneffekt, dass der Angreifer nun mit einer Wahrscheinlichkeit von 50 Prozent davon ausgehen kann, dass die Zielperson positiv getestet wurde – ein nicht zu unterschätzender Informationsgewinn [Zig+20].

Die zweite Angriffsmöglichkeit ergibt sich, wenn die Werte des sensiblen Attributs innerhalb einer Äquivalenzklasse zwar verschieden, aber semantisch ähnlich sind (sog. *Similarity Attack*). Die 3-diverse Tabelle 3.7 mit den sensiblen Attributen `Gehalt` und `Diagnose` verdeutlicht diesen Sachverhalt: Falls ein Angreifer über das Wissen verfügt, dass die Zielperson in der ersten Äquivalenzklasse zu finden ist, kann er daraus auf Magenprobleme schließen. Außerdem kann er ein im Vergleich zu den restlichen Einträgen relativ geringes Gehalt im Bereich von `3K` bis `5K` herauslesen [LLV07].

2 Die Entropie ist im Falle der Gleichverteilung der $p_{(E,s)}$ maximal.

PLZ	Geburtsjahr	Gehalt	Diagnose
761∗∗	199∗	3K	Magengeschwür
761∗∗	199∗	4K	Gastritis
761∗∗	199∗	5K	Magenkrebs
7622∗	198∗	6K	Gastritis
7622∗	198∗	11K	Grippe
7622∗	198∗	8K	Bronchitis
76∗∗∗	197∗	7K	Bronchitis
76∗∗∗	197∗	9K	Pneumonie
76∗∗∗	197∗	10K	Magenkrebs

Tabelle 3.7.: *Similarity Attack* [LLV07]

Um den Informationsgewinn eines Angreifers durch eine starke Abweichung der Attributverteilung zwischen einer Äquivalenzklasse und der gesamten Tabelle zu unterbinden, erdachten Li, Li und Venkatasubramanian [LLV07] das nachfolgend formalisierte Prinzip der t-Closeness.

Definition 3.3 (Prinzip der t-Closeness) Eine Äquivalenzklasse erfüllt t-Closeness, wenn der Abstand zwischen der Verteilung eines sensiblen Attributs innerhalb der Äquivalenzklasse und der Verteilung dieses Attributs in der gesamten Tabelle nicht größer als der Schwellwert t ist. Eine Tabelle besitzt t-Closeness, wenn alle Äquivalenzklassen t-Closeness erfüllen [Zig+20].

Mit t-Closeness kann also eine ungleichmäßige Attributverteilung (*Skewness Attack*) wirksam verhindert werden. Darüber hinaus kann die semantische Ähnlichkeit der Werte in der Äquivalenzklasse begrenzt werden, sofern sich nicht alle sensiblen Attributwerte in der Datenbank semantisch ähnlich sind. Darunter kann allerdings der Informationsgehalt – in Form der Korrelation zwischen quasi-identifizierenden und sensiblen Attributen – leiden. Der Parameter t erlaubt gewissermaßen einen *Trade-off* zwischen Nutzen und Datenschutz. Zu beachten ist, dass t-Closeness keine Garantien zur Vermeidung von *Identity Disclosure* gibt, weshalb eine zeitgleiche Verwendung von k-Anonymity sinnvoll scheint [LLV07].

Es existieren verschiedene Metriken, um den Abstand zwischen zwei Verteilungen A und B zu messen. Der Parameter t ergibt sich dann als Maximum der Abstände, wenn man ohne Beschränkung der Allgemeinheit für A die Verteilung innerhalb der Äquivalenzklassen und für B die globale Verteilung des sensiblen Attributs wählt. Um den Abstand zweier Werte hinsichtlich ihrer Semantik adäquat zu beurteilen, eignet sich die sogenannte *Earth Mover Distance* (EMD) in besonderem Maße. Intuitiv berechnet diese Metrik den minimalen Aufwand,

um die eine Verteilung in die andere zu überführen. Für ein numerisches Attribut wie `Gehalt`[3] lässt sich die EMD bei aufsteigender Sortierung der zu Grunde liegenden Attributwerte auf folgende Weise ermitteln, wobei $|A| = |B| = m$ [LLV07]:

$$\mathrm{EMD}(A, B) = \frac{1}{m-1} \sum_{i=1}^{i=m} \left| \sum_{j=1}^{i} (a_j - b_j) \right| .$$

Für kategoriale Attribute wie `Diagnose`, die im Gegensatz zu numerischen Attributen keine totale Ordnung besitzen, gibt es zwei verschiedene Verfahren zur Bestimmung der EMD. Die einfachere Variante geht davon aus, dass alle Attributwerte den gleichen Abstand voneinander haben, nämlich Eins.[4] Die EMD ist dann gerade die halbe Manhattan-Distanz [DSH19]:

$$\mathrm{EMD}(A, B) = \frac{1}{2} \sum_{i=1}^{m} |a_i - b_i| .$$

Bei der zweiten Variante basiert der Abstand zwischen zwei kategorialen Attributwerten auf der Anzahl an Generalisierungsschritten, die in der Generalisierungshierarchie (siehe Abschnitt 3.2.5.1) bis zu einem gemeinsamen Elternknoten ausgeführt werden müssen. In [LLV07] finden sich weitere Informationen zu dieser Variante. Eine exemplarische Berechnung des Parameters t für Tabelle 3.7 auf Basis der soeben dargestellten Definitionen kann in [DSH19] nachvollzogen werden. An dieser Stelle seien auch mit δ-Disclosure Privacy [BS08] und β-Likeness [CK12] zwei weitere Anonymitätsmaße erwähnt, die vor *Attribute Disclosure* schützen und wiederum gewisse Vor- und Nachteile im Vergleich zu t-Closeness aufweisen.

3.2.4.4. δ-Presence

Wie bereits erwähnt, kann die bloße An- oder Abwesenheit einer Person in einem Datensatz eine schützenswerte Erkenntnis darstellen, beispielsweise wenn in einer Datenbank Informationen einer stigmatisierten Erkrankung erfasst werden. Nergiz, Atzori und Clifton [NAC07] veröffentlichten mit δ-Presence ein Kriterium, das Schutz vor *Membership Disclosure* bietet. Voraussetzung für die Anwendbarkeit von δ-Presence ist, dass sowohl die Institution, welche eine Tabelle T für Forschungszwecke nutzbar machen möchte, als auch ein möglicher Angreifer Zugriff auf eine Tabelle P haben, die für alle Individuen sämtliche öffentlich verfügbaren Informationen umfasst. Das Ziel ist es nun, eine generalisierte Version T^* von T zu finden,

3 Im Fall von Tabelle 3.7 haben die aufsteigend sortierten Werte dieses Attributs die globale Verteilung $B_1 = \{1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9\}$, die erste Äquivalenzklasse besitzt die Verteilung $A_1 = \{1/3, 1/3, 1/3, 0, 0, 0, 0, 0, 0\}$.

4 In diesem Fall haben die sechs Attributwerte in Reihenfolge ihres ersten Auftretens die globale Verteilung $B_2 = \{1/9, 2/9, 2/9, 1/9, 2/9, 1/9\}$ und beispielhaft in der zweiten Äquivalenzklasse die Verteilung $A_2 = \{0, 1/3, 0, 1/3, 1/3, 0\}$.

welche die nachfolgende Definition erfüllt.

Definition 3.4 (δ-Presence) Die Tabelle P enthalte für jedes Individuum alle öffentlich verfügbaren Informationen. Sei T die zu anonymisierende Tabelle, wobei $T \subseteq P$. Dann erfüllt die generalisierte Version T^* von T δ-Presence für $\delta = (\delta_{min}, \delta_{max})$, wenn gilt [NC10]:

$$\delta_{min} \leq \Pr[t \in T \mid P, T^*] \leq \delta_{max}, \quad \forall t \in P.$$

Die Parameter δ_{min} und δ_{max} grenzen also den Bereich der Wahrscheinlichkeit ein, mit der sich jedes Individuum aus Angreifersicht im Forschungsdatensatz T befindet (oder eben auch nicht). Sei P^* die Tabelle, welche aus P durch diejenige Generalisierungsvorschrift hervorgeht, mit der auch T^* aus T erzeugt wurde. Dann lässt sich die Wahrscheinlichkeit für die Mitgliedschaft im Forschungsdatensatz für jedes Tupel t der Tabelle P wie folgt berechnen, wobei t^* die Generalisierung[5] von t in P^* bezeichne:

$$\Pr[t \in T \mid P, T^*] = \frac{|t^* \in T^*|}{|t^* \in P^*|}.$$

Es wird also das Verhältnis zwischen der Kardinalität einer jeden Äquivalenzklasse in der Tabelle T^* und der Kardinalität der entsprechenden Äquivalenzklasse in der Tabelle P^* gebildet [NC10].

Folgendes Beispiel soll die Vorgehensweise zur Bestimmung von δ_{min} und δ_{max} verdeutlichen: Die linke Tabelle in Abbildung 3.3 enthält die öffentlich bekannten Informationen, sie entspricht also P, der generalisierte Forschungsdatensatz T^* wird durch die rechte Tabelle repräsentiert. Die ersten sechs Tupel von P können in T^* der ersten, aus drei Vertretern bestehenden Äquivalenzklasse zugeordnet werden, daher beträgt die Wahrscheinlichkeit für die Mitgliedschaft im Forschungsdatensatz T für jedes dieser Tupel $3/6 = 1/2$. In Form der zweiten Äquivalenzklasse sind die letzten drei Einträge von P in T^* vertreten, somit beträgt die Wahrscheinlichkeit in diesem Fall jeweils $2/3$. Dementsprechend handelt es sich bei T^* um eine $(1/2, 2/3)$-präsente Tabelle [NAC07].

Allerdings ist die Verwendbarkeit von δ-Presence in der Praxis oftmals begrenzt, da im Allgemeinen keine Tabelle P zur Verfügung steht, in welcher alle öffentlich bekannten Daten zusammengefasst sind. Die in [NC10] vorgestellte Abwandlung c-Confident δ-Presence stellt realistischere Anforderungen: Hierbei wird davon ausgegangen, dass der veröffentlichenden Einrichtung nur begrenzte Informationen über die Bevölkerung in Form von Verteilungsfunktionen vorliegen [GLS14].

5 An dieser Stelle wird eine nicht überlappende Generalisierung vorausgesetzt.

P

Name	Alter	Land	PLZ
Alice	35	USA	47906
Bob	59	Kanada	47903
Chris	42	USA	47906
Dirk	18	Brasilien	47630
Eunice	22	Brasilien	47630
Frank	63	Peru	47633
Gail	33	Spanien	48973
Harry	47	Italien	48972
Iris	52	Frankreich	48970

T^*

Alter	Land	PLZ
*	Amerika	47***
*	Amerika	47***
*	Amerika	47***
*	Europa	48***
*	Europa	48***

Abbildung 3.3.: Beispiel für δ-Presence, wobei $\delta = \left(\frac{1}{2}, \frac{2}{3}\right)$ [NC10]

3.2.4.5. Differential Privacy

Differential Privacy (DP) wurde erstmals im Jahr 2006 von Cynthia Dwork et al. [Dwo+06] in einem Konferenzbeitrag vorgestellt und erfreut sich seither immer größerer Beliebtheit. Das Kriterium entstammt der Forschung im Bereich der statistischen Datenbanken. In diesem Setting werden hauptsächlich Aggregatfunktionen auf einzelne Spalten angewandt (siehe hierzu auch Abschnitt 3.2.1), der Zugriff auf einzelne Einträge ist nicht entscheidend [GLS14]. 1977 äußerte Dalenius [Dal77] das Desideratum, dass ein Angreifer, der Zugriff auf eine statistische Datenbank hat und mithilfe bestimmter Algorithmen Auswertungen durchführen kann, nichts über ein Individuum erfahren sollte, was er auch ohne diesen Datenbankzugriff erfahren hätte. Dwork zeigte in [Dwo06], dass wegen möglichen Hintergrundwissens kein Algorithmus die von Dalenius ersehnte semantische Sicherheit für statistische Datenbanken erreichen kann. Sie schlug mit DP einen neuartigen Ansatz vor, der statt einem absoluten Schutz der Privatsphäre auf relative Garantien, und das unabhängig vom Hintergrundwissen des Angreifers, setzt: Das Ergebnis einer Analyse darf sich durch Hinzufügen oder Entfernen eines einzelnen Eintrages nicht substanziell ändern, somit ist für jedes Individuum das Risiko, welches sich durch (Nicht-)Teilnahme an einer statistischen Datenbank ergibt, beschränkt [NW18]. Diese intuitive Betrachtungsweise lässt sich wie folgt formalisieren [Dwo08]:

Definition 3.5 (ε-Differential Privacy) Eine randomisierte Funktion $\mathcal{M}$ bietet ε-DP, wenn für alle Datensätze D_1 und D_2, die sich höchstens in einem Element unterscheiden, sowie für alle $S \subseteq \mathit{Bild}(\mathcal{M})$ gilt [Dwo06; MT07]:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D_2) \in S].$$

Im Gegensatz zu den zuvor genannten Ansätzen, bei denen die anonymisierte Tabelle

bestimmte syntaktische Bedingungen erfüllen musste, handelt es bei DP nicht um einen rein syntaktischen Begriff. Vielmehr wird von einem konkreten Mechanismus – in der Definition durch $\mathcal{M}$ repräsentiert – gefordert, dass jeder einzelne Eintrag das Ergebnis einer statistischen Auswertung nur geringfügig beeinflusst [CT13]. Dies gelingt, indem das exakte Ergebnis einer Anfrage mittels Perturbation (siehe Abschnitt 3.2.5.3) verrauscht wird. Da D_1 und D_2 in obiger Definition vertauscht werden können, impliziert der Parameter ε als Maß für die Stärke des zufälligen Rauschens sowohl eine obere als auch eine untere Grenze für die Abweichung, weshalb die im Kontext von DP eingesetzten Mechanismen nichtdeterministisch sind. Für die Wahl von ε werden in der Literatur häufig Werte zwischen 0,01 und $\ln 3$ genannt, wobei die Exaktheit der durch den Mechanismus ausgegebenen Ergebnisse mit steigendem ε zunimmt. So ist für $\varepsilon = 0{,}2$ beispielsweise eine Abweichung höchstens um den Faktor $e^{0{,}2} \approx 1{,}22$ erlaubt [Dwo08; GLS14].

Ein anderer Blickwinkel auf den Parameter ε ergibt sich aus der Kompositionseigenschaft von DP: Die sequentielle Ausführung von Mechanismen $\mathcal{M}_i$, die jeweils ε_i-DP erfüllen, gewährleistet ε-DP mit $\varepsilon = \sum_i \varepsilon_i$. Angesichts dessen bezeichnet man ε oftmals auch als *Privacy Budget*, welches man über die Wahl der ε_i auf die einzelnen Anfragen aufteilen kann. Auf diese Weise kann für jede Anfrage individuell zwischen Exaktheit des Ergebnisses und Datenschutz abgewogen werden. Sobald das *Privacy Budget* des Forschers aufgebraucht ist, muss ihm der Zugriff auf die Datenbank verwehrt werden [MT07; DE13].

Es gibt jedoch Szenarien, für die ε-DP möglicherweise zu strikt ist. Aus diesem Grund gibt es zahlreiche Variationen von DP, die bekannteste unter ihnen ist (ε, δ)-DP. Der Parameter δ gibt dabei an, dass ε-DP mindestens mit Wahrscheinlichkeit $1 - \delta$ erfüllt ist. Für δ sollten nur Werte infrage kommen, die vernachlässigbar in der Größe n der Tabelle sind.[6] Mathematisch ist (ε, δ)-DP folgendermaßen definiert [DE13; DR13]:

Definition 3.6 ((ε, δ)-Differential Privacy) Eine randomisierte Funktion $\mathcal{M}$ bietet (ε, δ)-DP, wenn für alle Datensätze D_1 und D_2, die sich höchstens in einem Element unterscheiden, sowie für alle $S \subseteq \mathit{Bild}(\mathcal{M})$ gilt [DR13; Wil+20]:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D_2) \in S] + \delta.$$

Mit *Local Differential Privacy* (LDP) steht ferner eine Variante bereit, bei der kein Vertrauen in die Einrichtung, welche die auszuwertenden Daten sammelt, nötig ist. Dies kann dadurch erreicht werden, dass die Teilnehmer ihre Daten zunächst selbst verrauschen – beispielsweise mit der im Abschnitt 3.2.5.3 beschriebenen Randomized-Response-Technik. Anschließend werden

6 Im Worst Case kann die Identität eines Individuums mit Wahrscheinlichkeit δ aufgedeckt werden, bei n Individuen summiert sich die Wahrscheinlichkeit für eine Re-Identifizierung auf $n \cdot \delta$.

die verrauschten Informationen an die Einrichtung weitergegeben. LDP wird in zunehmenden Maße von Technologiekonzernen zur Analyse von Nutzerdaten herangezogen: Apple nutzt das Verfahren, um neue Wörter, häufig verwendete Emojis oder speicherintensive Webseiten zu erkennen [DPT17], Google analysiert mit LDP Nutzungs- und Diagnosedaten seines Browsers Chrome [EPK14].

Darüber hinaus ist noch zu erwähnen, dass DP auch im nichtinteraktiven Szenario verwendet werden kann. Hierfür kommen *Synthetic Datasets* oder *Contingency Tables* zum Einsatz, eine detaillierte Erläuterung würde allerdings den Rahmen dieser Arbeit sprengen [Dwo08; DE13].

3.2.5. Anonymisierungsverfahren

Damit eine Tabelle im Bereich der nichtinteraktiven Forschungsdatennutzung die von einem Anonymitätsmaß geforderte syntaktische Struktur erfüllt, muss sie durch ein Anonymisierungsverfahren angepasst werden. Zwei im Folgenden vorgestellte Verfahren werden in diesem Zusammenhang besonders häufig verwendet: Generalisierung und Unterdrückung. Die anschließend betrachtete Perturbation ist dagegen für das Erreichen von DP fundamental. Auf die Beurteilung konkreter Implementierungen dieser Verfahren wird aus Platzgründen verzichtet. Mithilfe von Informationsmetriken, die in dieser Arbeit nicht erläutert werden, können verschiedene Anonymisierungen einer Tabelle verglichen werden, um einen angemessenen *Trade-off* zwischen dem Informationsgehalt der Daten und dem Schutz der Privatsphäre zu finden. Die wichtigsten Metriken werden in [Fun+10] und [Zig+20] vorgestellt.

3.2.5.1. Generalisierung

Durch Generalisierung (engl. *Generalization*) werden die Werte quasi-identifizierender Attribute durch allgemeinere Werte ersetzt. Indem unterschiedliche Werte auf denselben generalisierten Wert abgebildet werden, wird die Anzahl der Vertreter der jeweiligen Äquivalenzklasse erhöht. Die hierbei möglichen Attributwerte, die mit zunehmender Generalisierung immer ungenauer

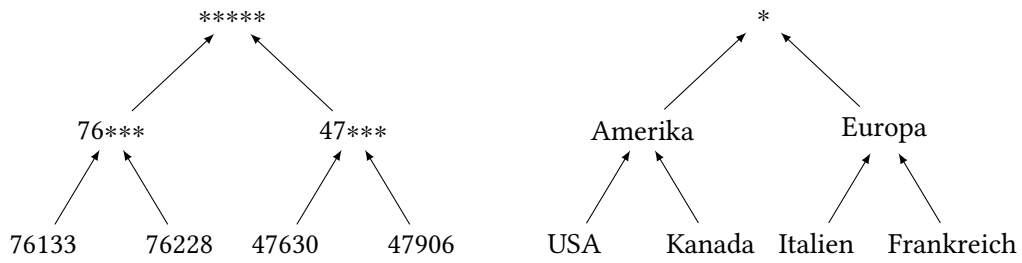

Abbildung 3.4.: Generalisierungshierarchien für PLZ und Land

werden, können dabei durch eine sogenannte Generalisierungshierarchie repräsentiert werden. Abbildung 3.4 zeigt beispielhaft zwei Generalisierungshierarchien für die Attribute PLZ und Land. Hierbei sei erwähnt, dass die automatische Erzeugung von Generalisierungshierarchien für kategoriale Attribute noch immer ein ungelöstes Problem darstellt [SS98; Zig+20].

Bei der Generalisierung können verschiedene Strategien unterschieden werden [Fun+10; Zig+20]:

Full-domain Generalization In diesem Fall werden alle Werte eines Attributs auf dieselbe Ebene in der Generalisierungshierarchie gebracht. Wird beispielsweise USA und Kanada zu Amerika generalisiert, so muss auch Italien und Frankreich durch Europa ersetzt werden.

Subtree Generalization Bei dieser Methode werden entweder alle oder keine Kindknoten eines inneren Knotens auf dieselbe Ebene generalisiert. Wird USA durch Amerika ersetzt, muss auch Kanada durch Amerika substituiert werden. Italien und Frankreich können allerdings unverändert bleiben.

Sibling Generalization Diese Vorgehensweise ähnelt der zuvor genannten. Hier können jedoch einige Kindknoten generalisiert werden, während andere unverändert bleiben. Beispielsweise könnte man alle Länder auf dem amerikanischen Kontinent mit Ausnahme von Kanada in Amerika generalisieren.

Cell Generalization Hier werden die Instanzen eines Attributwerts voneinander unabhängig manipuliert. So könnte USA bei einem Tupel zu Amerika generalisiert werden, die anderen Tupel mit dem Wert USA bleiben hingegen unverändert.

3.2.5.2. Unterdrückung

Unterdrückung (engl. *Suppression*) bezeichnet dagegen das Entfernen von Ausreißern aus dem Datensatz, die eine starke Generalisierung nach sich ziehen würden. In der Literatur werden drei verschiedene Ansätze zur Unterdrückung beschrieben [SS98; Fun+10]:

Record Suppression In diesem Fall wird der gesamte Eintrag entfernt.

Value Suppression Hierbei wird ein bestimmter Wert in der gesamten Tabelle unterdrückt.

Cell Suppression Bei dieser Vorgehensweise werden lediglich einige Instanzen eines gewissen Attributwerts eliminiert.

3.2.5.3. Perturbation

Auf dem Gebiet der DP existieren vielfältige Mechanismen, die zur Perturbation, also Verfremdung der tatsächlichen Ergebnisse, herangezogen werden. Für die nachfolgenden Betrachtungen muss zunächst die Sensitivität einer Anfrage eingeführt werden. Sie gibt an, wie abhängig

das Ergebnis von den Daten eines Individuums ist. Eine Anfrage kann dabei als Funktion f aufgefasst werden, die einer konkreten Datenbank D aus der Menge aller Datenbanken $\mathcal{D}$ einen reellen Vektor zuordnet [DR13].

Definition 3.7 (Sensitivität einer Anfrage) Eine Funktion $f\colon \mathcal{D} \mapsto \mathbb{R}^k$ besitzt die Sensitivität $\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$, wobei sich D_1 und D_2 höchstens in einem Element unterscheiden [Dwo08].

Stellvertretend wird nun der in der Praxis besonders häufig eingesetzte Laplace-Mechanismus erläutert. Die Laplace-Verteilung besitzt dabei in Abhängigkeit des Lageparameters μ und des Skalenparameters σ folgende Dichtefunktion:

$$g(x \mid \mu, \sigma) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}.$$

Der Laplace-Mechanismus verrauscht nun $f(X)$, das exakte Ergebnis einer Anfrage f für die Datenbank X, zu $f(X) + (Y_1, \ldots, Y_k)$, wobei $Y_i \sim \text{Laplace}(0, \Delta f/\varepsilon)$. Anschaulich wird also zu jeder Komponente des Ergebnisses ein Zufallswert aus der um Null zentrierten Laplace-Verteilung addiert. Aus Abbildung 3.5 geht hervor, wie die Wahl von ε die durchschnittliche Abweichung vom tatsächlichen Ergebnis beeinflusst: Je größer ε, desto spitzer ist die Verteilung. Damit steigt die Wahrscheinlichkeit, dass das Ergebnis nur in geringem Maße verfremdet wird. Im Grenzfall $\varepsilon \to 0$ ergibt sich dagegen eine flache Gerade, die sich dem Wert Null annähert – die vom Mechanismus zurückgegebenen Ergebnisse wären also zufällig und somit nutzlos [DR13; Zig+20].

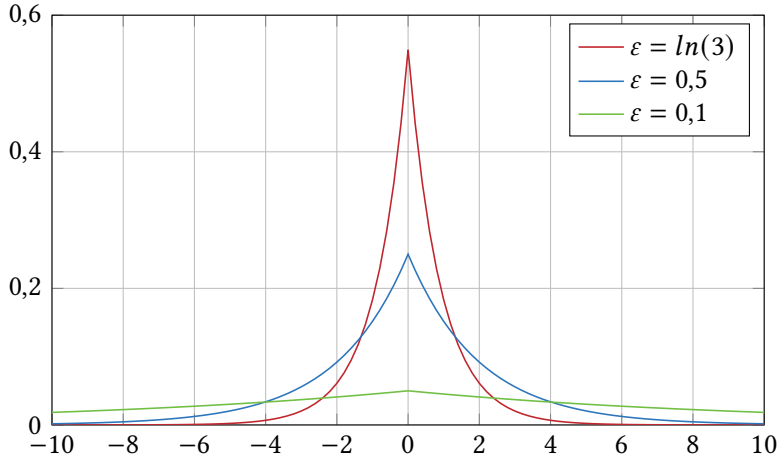

Abbildung 3.5.: Laplace-Verteilung mit $\mu = 0$ und $\sigma = \frac{1}{\varepsilon}$

Im Kontext der LDP findet häufig die Randomized-Response-Technik Anwendung, welche den Sozialwissenschaften entstammt. Hierbei wirft der Teilnehmer zunächst eine faire Münze (siehe auch Abbildung 3.6). Zeigt sie Kopf, so wird die Frage wahrheitsgemäß beantwortet. Zeigt sie Zahl, wird die Münze erneut geworfen und mit „Ja“ bei Kopf beziehungsweise „Nein“ bei Zahl geantwortet. Offensichtlich wird der Teilnehmer bei diesem Vorgehen die Frage mit einer Wahrscheinlichkeit von 75 Prozent wahrheitsgemäß beantworten. Die Randomized-Response-Technik erfüllt $\ln(3)$-DP, da

$$\frac{\Pr\left[\text{Antwort} = Ja \mid \text{Wahrheit} = Ja\right]}{\Pr\left[\text{Antwort} = Ja \mid \text{Wahrheit} = Nein\right]} = \frac{3/4}{1/4} = \frac{\Pr\left[\text{Antwort} = Nein \mid \text{Wahrheit} = Nein\right]}{\Pr\left[\text{Antwort} = Nein \mid \text{Wahrheit} = Ja\right]}.$$

Google verwendet diesen Mechanismus, um bestimmte Daten auszuwerten, die während der Nutzung des Browsers Chrome anfallen und als logisches Prädikat repräsentiert werden [DR13; EPK14].

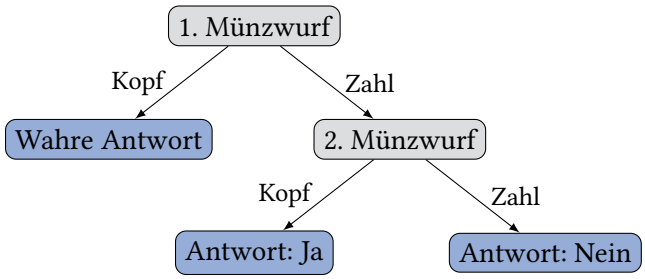

Abbildung 3.6.: Randomized-Response-Technik [EPK14]

Für einen Überblick über weitere Perturbationsoperationen können [Fun+10] und [Zig+20] konsultiert werden.

3.3. Homomorphe Verschlüsselung

Eine Alternative zur Anonymisierung stellen homomorphe Verschlüsselungsverfahren (engl. *Homomorphic Encryption Schemes*) dar. Derartige Verfahren erlauben es, beliebige Berechnungen direkt auf verschlüsselten Daten durchzuführen. Dabei ist die Anwendbarkeit nicht auf die Auswertung von Einträgen beschränkt, die in einer Datenbank gespeichert sind. Vielmehr ist eine Analyse beliebiger Datensätze beispielsweise durch neuronale Netze möglich. Dies ist insbesondere im Bereich des Cloud Computing interessant, wie aus Abbildung 3.7 hervorgeht: Der Auftraggeber verschlüsselt zunächst den Datensatz bei sich lokal und übergibt ihn an einen Dienstleister. Dieser führt die gewünschten Berechnungen auf den verschlüsselten Daten aus und gibt dem Auftraggeber das Ergebnis zurück, welches ebenfalls in verschlüsselter Form

vorliegt und vom Auftraggeber mit seinem geheimen Schlüssel entschlüsselt werden kann. Der Dienstleister sieht also zu keinem Zeitpunkt unverschlüsselte Informationen, weshalb ihm nicht vertraut werden muss. Aus diesem Grund erlauben homomorphe Verschlüsselungsverfahren eine datenschutzkonforme Auswertung medizinischer Daten mittels *High Performance Computing* (HPC) in der Cloud, womit ein Hochleistungsrechner vor Ort überflüssig wird [MHN15; Viz+19].

Abbildung 3.7.: Workflow einer Datenanalyse unter Einsatz von homomorpher Verschlüsselung [Viz+19]

Das Adjektiv „homomorph“ steht in diesem Zusammenhang für die algebraische Homomorphie-Eigenschaft, welche durch die Strukturerhaltung sinnvolle Berechnungen auf den Chiffraten ohne vorherige Entschlüsselung ermöglicht. Damit lässt sich ein homomorphes Verschlüsselungsverfahren für eine bestimmte Rechenoperation wie folgt definieren [FG07]:

Definition 3.8 (Homomorphes Verschlüsselungsverfahren) Sei M die Menge aller möglichen Nachrichten und C die Menge aller möglichen Chiffrate. Ein Verschlüsselungsverfahren mit Verschlüsselungsalgorithmus Enc und Entschlüsselungsalgorithmus Dec heißt homomorph bezüglich einer auf M definierten Operation $\star$, wenn eine Operation $\diamond$ auf C existiert, sodass gilt [FG07; Aca+18]:

$$\textsc{Dec}(\textsc{Enc}(m_1) \diamond \textsc{Enc}(m_2)) = \textsc{Dec}(\textsc{Enc}(m_1 \star m_2)) = m_1 \star m_2, \quad \forall m_1, m_2 \in M.$$

Ein Verschlüsselungsverfahren heißt additiv (beziehungsweise multiplikativ) homomorph, wenn die auf C definierte Operation der Addition (beziehungsweise Multiplikation) auf M entspricht. Addition (XOR) und Multiplikation (AND) bilden ein vollständiges Operatorensystem, weshalb mit ihnen beliebige Funktionen ausgewertet werden können [Aca+18]. Bereits 1978 entdeckten Rivest, Adleman und Dertouzos [RAD78], dass Textbook-RSA multiplikativ homomorph ist, was sie als *Privacy Homomorphism* bezeichneten. Damit handelt es sich bei Textbook-RSA um ein sogenanntes *Partially Homomorphic Encryption Scheme*, bei welchem eine Operation – entweder Addition oder Multiplikation – beliebig oft verwendet kann [Aca+18].

Demgegenüber veröffentlichte Gentry [Gen09] im Jahr 2009 erstmals ein *Fully Homomorphic Encryption Scheme*, welches die Auswertung von beliebig vielen Additionen und Multiplikatio-

nen erlaubt. Allerdings ist die praktische Relevanz aller bislang bekannten Verfahren, welche die zuvor genannte Eigenschaft erfüllen, begrenzt, da die Chiffrate während der Berechnung regelmäßig mittels einer zeitintensiven Prozedur (sog. *Bootstrapping*) umgeschlüsselt werden müssen. Aus praktischer Sicht sind daher *Somewhat Homomorphic Encryption Schemes* interessanter, bei denen eine wesentlich effizientere Auswertung für eine begrenzte Anzahl von Additionen und Multiplikationen möglich ist [Aca+18].

Um den Rahmen dieser Arbeit nicht zu sprengen, werden homomorphe Verschlüsselungsverfahren im Folgenden nicht weiter betrachtet.

4. Prototypische Forschungsschnittstelle

In diesem Kapitel wird die prototypische Forschungsschnittstelle namens *PRIvacy cOmpliant Research Interface* (PRIORI) vorgestellt, die im Rahmen dieser Bachelorarbeit entwickelt wurde.

4.1. Systemkontext

PRIORI kann prinzipiell überall dort verwendet werden, wo besonders schützenswerte Informationen Außenstehenden datenschutzgerecht zugänglich gemacht werden sollen. Im Folgenden wird von einem Krankenhaus als Einsatzort ausgegangen, das seine Patientendaten in einem IT-System speichert (siehe Abbildung 4.1). Ein Medizininformatiker ruft den Datensatz, welcher der Forschung zur Verfügung gestellt werden soll, im IT-System ab und lädt diesen auf PRIORI hoch. Soll einem Forscher eine anonymisierte Version des Datensatzes im Rahmen eines Forschungsprojekts bereitgestellt werden, so übermittelt der Medizininformatiker die gewünschten Parameter der gewählten syntaktischen Anonymitätsmaße an PRIORI. Anschließend transformiert der Prototyp den Datensatz mittels Generalisierung und Unterdrückung so, dass die Anonymitätskriterien erfüllt sind. Sobald dieser Prozess abgeschlossen ist, kann der

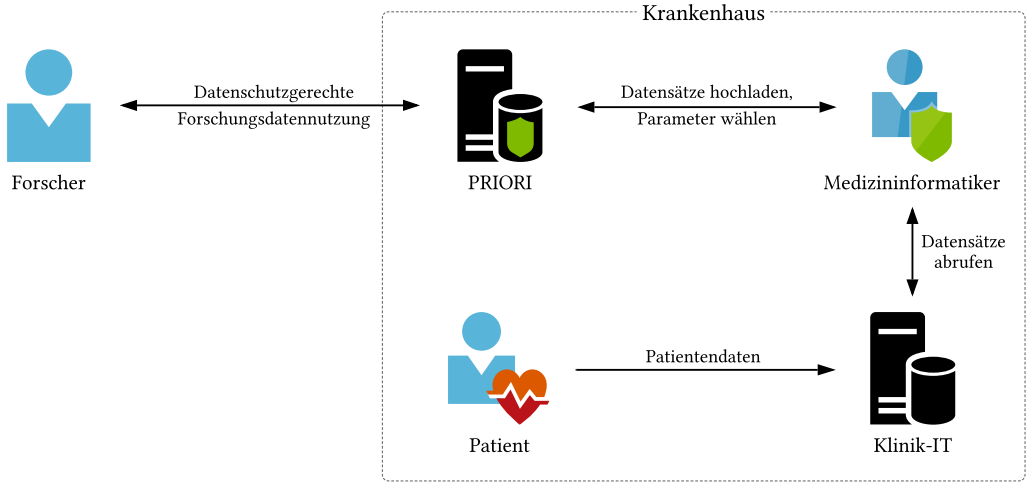

Abbildung 4.1.: Systemkontext von PRIORI am Beispiel eines Krankenhauses

Forscher den anonymisierten Datensatz herunterladen und auswerten. Alternativ unterstützt PRIORI auch das interaktive Szenario, bei welchem vom Forscher lediglich einfache statistische Auswertungen durch SQL-Anfragen durchgeführt werden können. Hierbei werden die Ergebnisse gemäß der Definition von DP (siehe Abschnitt 3.2.4.5) verrauscht.

In diesem Szenario wird davon ausgegangen, dass der Mitarbeiter im Krankenhaus (als „Medizininformatiker“) sowohl über das nötige medizinische als auch technische Wissen verfügt, um die Parameter für das jeweilige (Forschungs-)Projekt zweckmäßig zu wählen.

4.2. Architektur

PRIORI setzt auf das Architekturmuster *Representational State Transfer* (REST). Dieser Architekturstil wurde von Roy Thomas Fielding, der an der Entwicklung zahlreicher Standards wie dem *Hypertext Transfer Protocol* (HTTP) beteiligt war, im Rahmen seiner Dissertation [Fie00] für Webservices entwickelt. Ein auf Basis des REST-Architekturstils entwickelter Dienst zeichnet sich durch folgende Eigenschaften aus:

Client–Server Durch ein Client–Server-Modell wird die Datenhaltung von der Benutzerschnittstelle getrennt (*Separation of Concerns*). Hierdurch können die Komponenten voneinander unabhängig weiterentwickelt oder ersetzt werden, solange die Schnittstelle unverändert bleibt [Mas11].

Zustandslosigkeit Jede Anfrage des Clients enthält alle Informationen, welche der Server zur Verarbeitung benötigt. Die zustandslose Kommunikation stellt zwar einen gewissen Overhead dar, verbessert aber die Zuverlässigkeit und Skalierbarkeit [Fie00; LBÜ15].

Einheitliche Schnittstelle Dieses Prinzip fordert zum einen, dass jede Ressource durch einen *Uniform Resource Identifier* (URI) adressiert werden kann. Zum anderen müssen REST-Nachrichten selbstbeschreibend sein, was sich durch die Semantik der verschiedenen HTTP-Methoden – `GET`, `POST`, `PUT`, `PATCH`, `DELETE` und `OPTIONS` – erreichen lässt [Mas11].

Caching Clients und Zwischensysteme zwischen Client und Server können Antworten zwischenspeichern und bei einer identischen Anfrage zur Steigerung der Effizienz wiederverwenden. Hierzu gibt der Server bei jeder Antwort an, ob diese zwischengespeichert werden darf [Fie00].

4.3. Entwurf

Für den Prototyp wurden drei Ressourcen identifiziert: Nutzer, Datensätze und Projekte. Zwischen den beiden letztgenannten besteht dabei eine 1:n-Beziehung, ein Datensatz kann also im

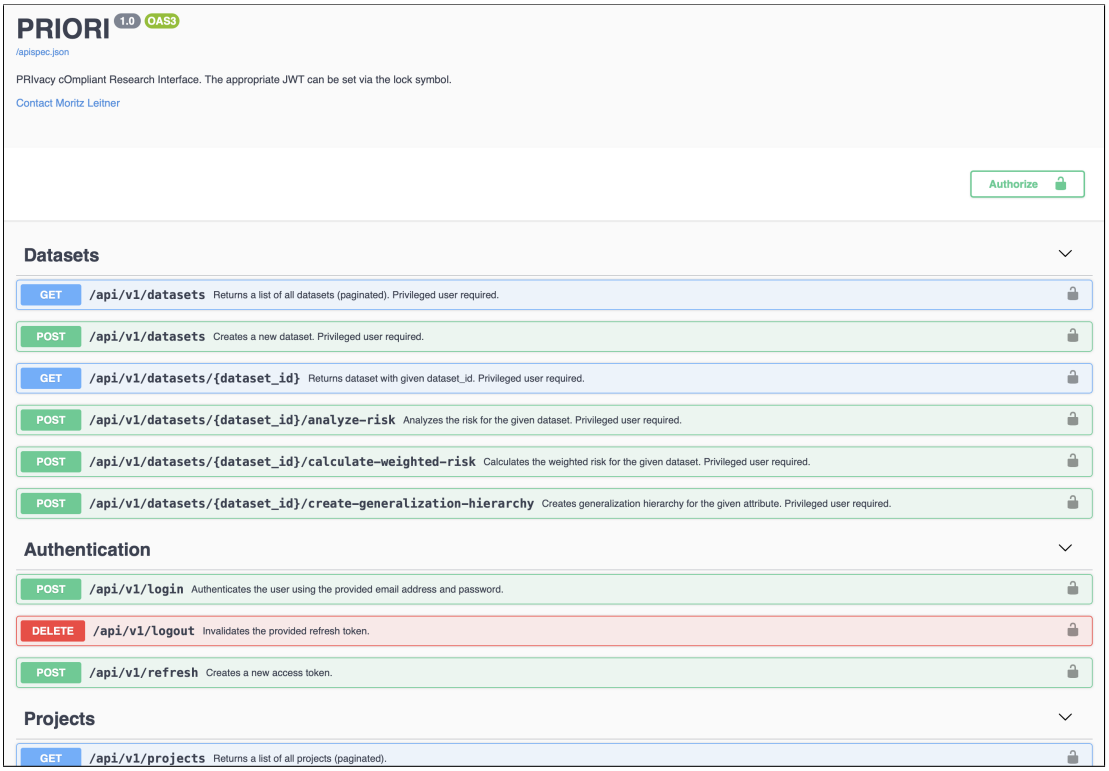

Abbildung 4.2.: Visualisierung der OpenAPI-Spezifikation durch Swagger UI (Ausschnitt)

Rahmen von mehreren Projekten verschiedenen Nutzern (mit unterschiedlichen Parametern der verwendeten PETs) zugänglich gemacht werden. Des Weiteren wird zwischen privilegierten und nicht privilegierten Nutzern unterschieden. Erstere können neue Nutzer anlegen, Datensätze hochladen sowie Projekte erstellen und löschen. Demgegenüber können nicht privilegierte Nutzer nur den für sie im Rahmen eines Projekts freigegebenen, anonymisierten Datensatz abrufen oder im interaktiven Szenario durch SQL-Anfragen statistisch auswerten.

Das *Application Programming Interface* (API) und die zugehörigen Routen wurden nach den Richtlinien in [Mas11] entworfen. Die ausführliche OpenAPI-Spezifikation kann mittels Swagger UI im Browser betrachtet werden (siehe Abbildung 4.2). Für weitere Informationen hierzu sei auf das *Repository*[1] von PRIORI verwiesen. Nachfolgend werden die wichtigsten API-Endpunkte für das Krankenhaus-Szenario vorgestellt, Beispiele für die jeweiligen JSON-Objekte finden sich im Anhang.

Zunächst muss sich der Medizininformatiker gegenüber PRIORI authentisieren. Hierzu übermittelt er seine Log-in-Daten an den Log-in-Endpunkt. Anschließend werden die Log-in-Daten von PRIORI überprüft, bei erfolgreicher Authentifizierung werden zwei *JSON Web*

1 `https://gitlab.com/leitmori/bachelor-thesis-code`

Token zurückgegeben:

POST /api/v1/login

Request Body Das erwartete JSON-Objekt hat folgendes Format:

Attribut	Typ	Beschreibung
`email_address`	String	E-Mail-Adresse des Nutzers.
`password`	String	Passwort des Nutzers.

Mögliche Fehler

Unauthorized: Der Anfragende hat ungültige Log-in-Daten übermittelt.

Antwort Statuscode, JSON-Objekt mit *Access-Token* und *Refresh-Token*.

Das *Access-Token*, welches defaultmäßig eine Gültigkeit von fünfzehn Minuten besitzt, muss in jeder weiteren Anfrage im HTTP-Header mitgeschickt werden. Mit dem *Refresh-Token*, das mit sieben Tagen eine deutlich längere Gültigkeit besitzt, kann bei Bedarf ein neues *Access-Token* angefordert werden, was hier aus Platzgründen weggelassen wird.

Der Medizininformatiker kann nun für den Forscher ein Benutzerkonto anlegen, indem er die Stammdaten des Forschers an den entsprechenden API-Endpunkt sendet:

POST /api/v1/users

HTTP-Header *Access-Token* im Bearer-Schema.

Request Body Das erwartete JSON-Objekt hat folgendes Format:

Attribut	Typ	Beschreibung
`email_address`	String	E-Mail-Adresse des Nutzers.
`institution`	String	Institution, welcher der Nutzer angehört.
`name`	String	Name des Nutzers.
`password`	String	Passwort des Nutzers.
`privileged_status`	Boolean	Gibt an, ob der Nutzer privilegiert ist oder nicht.

Mögliche Fehler

Bad Request: Das übergebene JSON-Objekt ist ungültig.

Unauthorized: Das *Access-Token* im HTTP-Header fehlt.

Forbidden: Der Anfragende ist nicht privilegiert.

Antwort Statuscode, erstellter Nutzer als JSON-Objekt, URI des erstellten Nutzers im HTTP-Location-Header.

Nun befindet sich der Forscher auch in der Liste aller Nutzer, die folgender Endpunkt zurückliefert:

GET /api/v1/users?*limit=20&page=1*

HTTP-Header *Access-Token* im Bearer-Schema.

Query-Parameter Der Query-String kann folgende optionale Parameter enthalten:

Parameter	Typ	Beschreibung
`limit`	Integer	Die maximale Anzahl an Nutzern (defaultmäßig: 20), die zurückgegeben werden soll.
`page`	Integer	Die angefragte Seite (defaultmäßig: 1).

Mögliche Fehler

Unauthorized: Das *Access-Token* im HTTP-Header fehlt.

Forbidden: Der Anfragende ist nicht privilegiert.

Antwort Statuscode, Nutzer als JSON-Objekt.

Bevor dem Forscher im Rahmen eines Projekts Zugriff auf einen (anonymisierten) Datensatz gewährt werden kann, muss der Medizininformatiker den Datensatz zunächst durch einen *Multipart Request* an den entsprechenden Endpunkt hochladen:

POST /api/v1/datasets

HTTP-Header *Access-Token* im Bearer-Schema.

Request Body Im *Request Body* müssen sowohl der Datensatz als CSV-Datei sowie zugehörige Metainformationen als JSON-Objekt übermittelt werden. Das JSON-Objekt hat hierbei folgendes Format:

Attribut	Typ	Beschreibung
`name`	String	Name des Datensatzes.
`description`	String	Beschreibung des Datensatzes.

Mögliche Fehler

Bad Request: Das übergebene JSON-Objekt ist ungültig oder die CSV-Datei fehlt.

Unauthorized: Das *Access-Token* im HTTP-Header fehlt.

Forbidden: Der Anfragende ist nicht privilegiert.

Antwort Statuscode, erstellter Datensatz als JSON-Objekt, URI des erstellten Datensatzes im HTTP-Location-Header.

Hierauf wird der Datensatz in die Liste aller Datensätze aufgenommen, welche über folgenden Endpunkt abgerufen werden kann:

GET /api/v1/datasets?*limit=20&page=1*

HTTP-Header *Access-Token* im Bearer-Schema.

Query-Parameter Der Query-String kann folgende optionale Parameter enthalten:

Parameter	Typ	Beschreibung
`limit`	Integer	Die maximale Anzahl an Datensätzen (defaultmäßig: 20), die zurückgegeben werden soll.
page	Integer	Die angefragte Seite (defaultmäßig: 1).

Mögliche Fehler

Unauthorized: Das *Access-Token* im HTTP-Header fehlt.

Forbidden: Der Anfragende ist nicht privilegiert.

Antwort Statuscode, Datensätze als JSON-Objekt.

Soll der Datensatz anonymisiert werden, muss für quasi-identifizierende Attribute eine Generalisierungshierarchie angegeben werden. PRIORI kann Generalisierungshierarchien mithilfe zweier verschiedener Transformationen generieren:

POST /api/v1/datasets/<*dataset_id*>/create-generalization-hierarchy

HTTP-Header *Access-Token* im Bearer-Schema.

Path-Parameter ID des Datensatzes.

Request Body Das erwartete JSON-Objekt hat folgendes Format:

Attribut	Typ	Beschreibung
`attribute_name`	String	Name des Attributs, für welches die Generalisierungshierarchie erstellt werden soll.
`transformation_type`	String	Art der Transformation, welche auf die Attributwerte angewandt werden soll (entweder `redaction` oder `interval`, siehe hierzu Beispiele im Anhang).
`intervals`	Array	Array von Intervallen (im Falle von `transformation_type = interval`).

`redaction_order`	String	Reihenfolge, in welcher die Attributwerte transformiert werden sollen (im Falle von `transformation_type = redaction`).

Mögliche Fehler
Bad Request: Das übergebene JSON-Objekt ist ungültig.
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.
Forbidden: Der Anfragende ist nicht privilegiert.
Internal Server Error: Das Bearbeiten der Anfrage war nicht erfolgreich.

Antwort Statuscode, erstellte Generalisierungshierarchie als JSON-Objekt.

Um für die Anonymitätskriterien geeignete Parameter und damit einen angemessenen *Trade-off* zwischen dem Informationsgehalt der Daten und dem Schutz der Privatsphäre zu bestimmen, kann der Medizininformatiker unter dem folgenden Endpunkt eine ausführliche Risikobewertung abrufen:

POST /api/v1/datasets/<*dataset_id*>/analyze-risk

HTTP-Header *Access-Token* im Bearer-Schema.

Path-Parameter ID des Datensatzes.

Request Body Das erwartete JSON-Objekt hat folgendes Format:

Attribut	Typ	Beschreibung
`attributes`	Array	Array, das für jedes Attribut den Typ (`identifying`, `quasi-identifying`, `sensitive` oder `insensitive`) und im Falle eines quasi-identifizierenden Attributs die zugehörige Generalisierungshierarchie enthält.
`privacy_models`	Array	Array, das die Anonymitätskriterien mit den entsprechenden Parametern enthält, die bei der Anonymisierung des Datensatzes und der anschließenden Risikobewertung verwendet werden sollen.
`suppression_limit`	Float	Maximaler Anteil an Einträgen, die durch Unterdrückung entfernt werden.

Mögliche Fehler
Bad Request: Das übergebene JSON-Objekt ist ungültig.
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.
Forbidden: Der Anfragende ist nicht privilegiert.
Internal Server Error: Das Bearbeiten der Anfrage war nicht erfolgreich.

Antwort Statuscode, Risikobewertung als JSON-Objekt.

Das Krankenhaus hat darüber hinaus die Möglichkeit, die in der Risikobewertung enthaltenen Maße im Rahmen einer im Konfigurationsmodul (`config.py`) von PRIORI hinterlegten *Policy* zu gewichten. Der nachstehende Endpunkt berechnet das gewichtete Risiko:

POST /api/v1/datasets/<*dataset_id*>/calculate-weighted-risk

HTTP-Header *Access-Token* im Bearer-Schema.

Path-Parameter ID des Datensatzes.

Request Body Das erwartete JSON-Objekt hat folgendes Format:

Attribut	Typ	Beschreibung
`attributes`	Array	Array, das für jedes Attribut den Typ (`identifying`, `quasi-identifying`, `sensitive` oder `insensitive`) und im Falle eines quasi-identifizierenden Attributs die zugehörige Generalisierungshierarchie enthält.
`privacy_models`	Array	Array, das die Anonymitätskriterien mit den entsprechenden Parametern enthält, die bei der Anonymisierung des Datensatzes und der Risikobewertung verwendet werden sollen.
`suppression_limit`	Float	Maximaler Anteil an Einträgen, die durch Unterdrückung entfernt werden.

Mögliche Fehler
Bad Request: Das übergebene JSON-Objekt ist ungültig.
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.
Forbidden: Der Anfragende ist nicht privilegiert.
Internal Server Error: Das Bearbeiten der Anfrage war nicht erfolgreich.

Antwort Statuscode, gewichtetes Risiko als JSON-Objekt.

Im Rahmen eines Forschungsprojekts kann dem Forscher Zugriff auf einen (anonymisierten) Datensatz gewährt werden. Hierzu schickt der Medizininformatiker eine Anfrage mit den zu verwendenden PETs und den zugehörigen Metadaten an folgenden Endpunkt, woraufhin der spezifizierte Datensatz im nichtinteraktiven Szenario anonymisiert wird:

POST /api/v1/projects

HTTP-Header *Access-Token* im Bearer-Schema.

Request Body Das erwartete JSON-Objekt hat folgendes Format:

Attribut	Typ	Beschreibung
`name`	String	Name des Projekts.
`description`	String	Beschreibung des Projekts.
`researcher_id`	Integer	ID des Forschers, für welchen das Projekt freigegeben werden soll.
`dataset_id`	Integer	ID des im Projekt verwendeten Datensatzes.
`config`	JSON	PETs mit entsprechenden Parametern, die beim Projekt zum Einsatz kommen.

Mögliche Fehler

Bad Request: Das übergebene JSON-Objekt ist ungültig.
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.
Forbidden: Der Anfragende ist nicht privilegiert.

Antwort Statuscode, erstelltes Projekt als JSON-Objekt, URI des erstellten Projekts im HTTP-Location-Header.

Sodann erscheint das Projekt in der Projektübersicht, die im Falle des Forschers alle mit ihm geteilten Projekte enthält. Als privilegierter Nutzer erhält der Medizininformatiker unter dem nachstehenden Endpunkt dagegen alle Projekte, die bislang erstellt wurden:

GET /api/v1/projects*?limit=20&page=1*

HTTP-Header *Access-Token* im Bearer-Schema.

Query-Parameter Der Query-String kann folgende optionale Parameter enthalten:

Parameter	Typ	Beschreibung
`limit`	Integer	Die maximale Anzahl an Projekten (default-mäßig: 20), die zurückgegeben werden soll.

page	Integer	Die angefragte Seite (defaultmäßig: 1).

Mögliche Fehler
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.

Antwort Statuscode und Projekte als JSON-Objekt.

Wurde für das Projekt das nichtinteraktive Szenario gewählt, kann der Forscher den anonymisierten Datensatz als CSV-Datei herunterladen:

GET /api/v1/projects/<*project_id*>/csv

HTTP-Header *Access-Token* im Bearer-Schema.

Path-Parameter ID des Projekts.

Mögliche Fehler
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.
Forbidden: Der Anfragende ist nicht privilegiert und das angefragte Projekt wurde nicht für ihn freigegeben.

Antwort Statuscode, anonymisierter Datensatz als CSV-Datei.

Im Rahmen des interaktiven Szenarios steht dem Forscher für statistische Auswertungen dagegen folgender Endpunkt zur Verfügung:

POST /api/v1/projects/<*project_id*>/analyze

HTTP-Header *Access-Token* im Bearer-Schema.

Path-Parameter ID des Projekts.

Request Body Das JSON-Objekt enthält unter dem *Key* `query` die SQL-Anfrage, welche ausgeführt werden soll. Hierbei werden die Aggregatfunktionen `COUNT`, `SUM`, `AVG`, `VARIANCE`, `STDDEV` sowie `GROUP BY` unterstützt.

Mögliche Fehler
Bad Request: Das übergebene JSON-Objekt ist ungültig.
Unauthorized: Das *Access-Token* im HTTP-Header fehlt.
Forbidden: Der Anfragende ist nicht privilegiert und das angefragte Projekt wurde nicht für ihn freigegeben.

Antwort Statuscode, unter DP verrauschtes Ergebnis der Anfrage als JSON-Objekt.

4.4. Implementierung

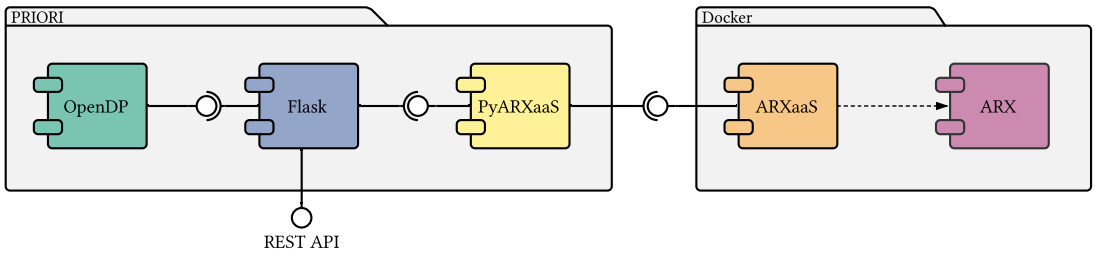

Abbildung 4.3.: UML-Komponentendiagramm des Prototyps

Wie aus Abbildung 4.3 hervorgeht, wurde für die Implementierung des REST-API von PRIORI das Python-Webframework *Flask*[2] verwendet. Alle von außen eingehenden Anfragen werden somit von der Flask-Anwendung verarbeitet und beantwortet. Als Microframework beschränkt sich *Flask* auf die absolut notwendige Kernfunktionalität. Eine große Anzahl bestehender Erweiterungen kann flexibel nachgeladen werden, um den Funktionsumfang auszuweiten. So kommt *SQLAlchemy*[3] beispielsweise für das *Object-Relational Mapping* (ORM) zum Einsatz, während für die Serialisierung und Deserialisierung auf *marshmallow*[4] zurückgegriffen wird. Die Verarbeitung der Datensätze erfolgt mithilfe der Bibliothek *pandas*[5] als `DataFrame`.

Die beiden wichtigsten Komponenten, mit denen die Flask-Anwendung interagiert, sind die Open-Source-Lösungen *ARX* und *OpenDP* (siehe hierzu auch Abschnitt 2.3). *OpenDP* erlaubt es, SQL-Anfragen auf einem `DataFrame` auszuführen, wobei die Ergebnisse gemäß der Definition von DP verfremdet werden. Zur Anonymisierung und Risikobewertung von Datensätzen sowie zur Erstellung von Generalisierungshierarchien wird – zumindest indirekt – *ARX* eingesetzt. Da eine unmittelbare Nutzung des in Java geschriebenen *ARX* in Python-Code nicht anzuraten ist, werden in der Flask-Anwendung die von *PyARXaaS*[6] bereitgestellten Python-Bindings verwendet. Diese Bibliothek macht wiederum mit *ARX as a Service* (ARXaaS)[7] von einer per Docker-Container bereitgestellten ARX-Instanz Gebrauch.

PRIORI steht ebenfalls als Docker-Image zur Verfügung, um mögliche Inkompatibilitäten mit der gerätespezifischen Python-Umgebung zu vermeiden. In einer Produktivumgebung sollte die Flask-Anwendung von *Gunicorn*[8] hinter einem Proxy-Server wie *NGINX* ausgeliefert

2 `https://flask.palletsprojects.com` (besucht am 01. 10. 2020)
3 `https://www.sqlalchemy.org` (besucht am 01. 10. 2020)
4 `https://github.com/marshmallow-code/marshmallow` (besucht am 01. 10. 2020)
5 `https://pandas.pydata.org` (besucht am 01. 10. 2020)
6 `https://github.com/navikt/pyarxaas` (besucht am 02. 10. 2020)
7 `https://github.com/navikt/ARXaaS` (besucht am 02. 10. 2020)
8 `https://gunicorn.org` (besucht am 02. 10. 2020)

werden. Eine Anleitung zur lokalen Installation sowie zur Konfiguration von PRIORI ist im *Repository* zu finden.

5. Evaluation

Im Folgenden wird PRIORI, die prototypische Forschungsschnittstelle zur Anonymisierung und statistischen Auswertung von Datensätzen, evaluiert. Hierbei werden zunächst die Performanz und die Auswirkungen unterschiedlicher Parameter für ausgewählte PETs untersucht. Anschließend werden die Möglichkeiten und Einschränkungen des gewählten Ansatzes diskutiert.

Die nachfolgenden Auswertungen wurden auf einem MacBook Pro 2017 mit einer Intel® Core™ i7-7820HQ CPU und 16 GiB LPDDR3-RAM unter Python 3.8.5 durchgeführt. Die zur Analyse verwendeten Datensätze wurden mit dem Testdatengenerator *Mimesis*[1] erzeugt. Hierfür existieren auch speziell für den medizinischen Bereich entwickelte Werkzeuge wie der *Synthea™ Patient Generator*,[2] welcher ausreichend große Datensätze allerdings nicht mit akzeptabler Geschwindigkeit liefern konnte. Bei den von *Synthea™* generierten Datensätzen werden außerdem einzelne Attribute absichtlich nicht bei jedem Patienten „ausgefüllt“, weshalb für die Anwendbarkeit der PETs zusätzliche Vorverarbeitungsschritte notwendig wären. Um realistische Ergebnisse für den vorgesehenen Einsatzkontext von PRIORI zu erhalten, wurden im Testskript nicht die jeweiligen Methoden direkt aufgerufen. Vielmehr wurden mithilfe des `werkzeug.test.Client`[3] HTTP-Anfragen an die entsprechenden API-Endpunkte gesendet und die Antwortzeiten gemessen.

Die für das nichtinteraktive Szenario generierten Datensätze bestehen aus den fünf Attributen Sozialversicherungsnummer, Postleitzahl, Alter, Geschlecht und Blutgruppe. Somit gilt für die Menge der quasi-identifizierenden Attribute QI = {PLZ, Alter, Geschlecht}, während die Blutgruppe das sensible Attribut ist. Im Rahmen des interaktiven Szenarios werden ebenfalls Beispieldaten mit den genannten Attributen genutzt, lediglich die Sozialversicherungsnummer wird als im Anwendungsfall ungeeigneter direkter Identifikator entfernt.

5.1. Performanz

Auch wenn das interaktive und das nichtinteraktive Szenario prinzipbedingt unterschiedliche Anwendungsbereiche und damit auch grundlegend verschiedene Ausgaben aufweisen, so

1 `https://github.com/lk-geimfari/mimesis` (besucht am 05. 10. 2020)
2 `https://github.com/synthetichealth/synthea` (besucht am 05. 10. 2020)
3 `https://werkzeug.palletsprojects.com/test/#werkzeug.test.Client` (besucht am 06. 10. 2020)

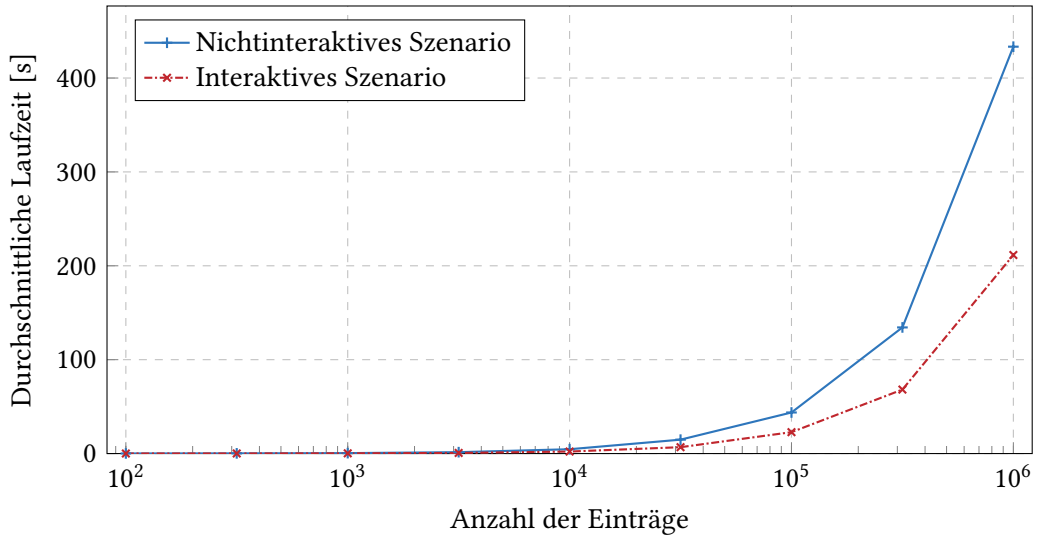

Abbildung 5.1.: Durchschnittliche Ausführungszeiten bei steigender Datensatzgröße

unterscheiden sie sich aus Sicht von PRIORI nur geringfügig: In beiden Fällen wird der als CSV-Datei vorliegende Datensatz in einen `pandas.DataFrame` eingelesen und entweder zur Ausführung der übermittelten SQL-Anfrage an *OpenDP* oder zur Anonymisierung an *PyARXaaS* übergeben. Aus diesem Grund sollte für beide Varianten das Laufzeitverhalten bei steigenden Datenmengen mit beispielhaften Parameterwerten untersucht und gegenübergestellt werden. Im nichtinteraktiven Szenario wurde die benötigte Dauer für das Anlegen eines neues Projekts und der damit verbundenen Anonymisierung des zugrunde liegenden Datensatzes ermittelt. Zur Vermeidung einer Re-Identifizierung mithilfe der quasi-identifizierenden Attribute wurde 5-Anonymity verwendet. Das sensible Attribut wurde zusätzlich mit Distinct 2-Diversity geschützt, in jeder Äquivalenzklasse müssen sich also zwei verschiedene Blutgruppen befinden. Im Kontext des interaktiven Szenarios wurde unter Gewährleistung von DP mit den Parametern $\varepsilon = 3$ und $\delta = 10^{-12}$ die Anfrage `SELECT blood_type, AVG(age) FROM PRIORI GROUP BY blood_type` ausgeführt.

Schaubild 5.1 zeigt für ausgewählte Datensatzgrößen die durchschnittliche Laufzeit aus fünf Durchläufen in halblogarithmischer Darstellung, wobei für jeden Durchlauf und jede Anzahl von Einträgen ein neuer Datensatz generiert wurde. Es ist ersichtlich, dass die Laufzeit bis zur Größenordnung von wenigen Tausend Einträgen nahezu konstant bleibt. Eine genauere Betrachtung des darauffolgenden Bereichs zwischen 25 000 und 250 000 Einträgen, für den die minimale, maximale und durchschnittliche Laufzeit bei zehn Durchläufen gesondert ausgewertet wurde (siehe Abbildung 5.2), lässt einen linearen Anstieg der Laufzeit erkennen. Zusätzlich benötigt die Anonymisierung der Datensätze etwa doppelt so lange wie die Ausführung der

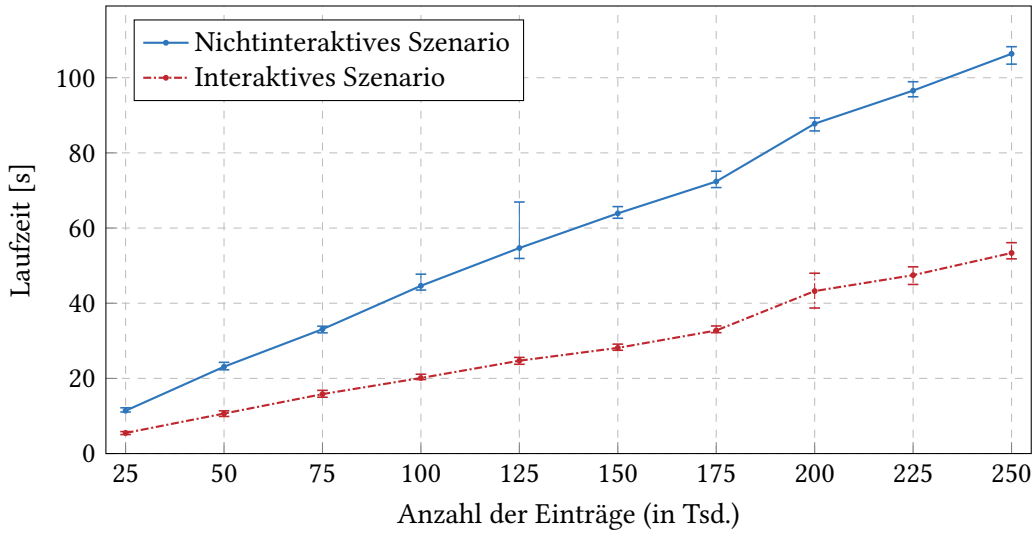

Abbildung 5.2.: Ausführungszeiten bei steigender Datensatzgröße

SQL-Anfrage. Dieser Effekt kann auf die erforderliche Kommunikation zwischen *PyARXaaS* und ARXaaS, die auf Basis von HTTP erfolgt, zurückgeführt werden.

Um die Frage zu beantworten, inwiefern die Wahl des Parameters k bei k-Anonymity die Ausführungszeit beeinflusst, wurde zunächst ein Datensatz mit 10 000 Einträgen generiert. Anschließend wurden mit diesem Datensatz (`non_interactive.csv` im *Repository*) neue Projekte mit unterschiedlichen Werten für k erstellt und die Laufzeit gemessen. Die Blutgruppe wurde jeweils mit Distinct 2-Diversity geschützt. Aus Abbildung 5.3 geht hervor, dass die

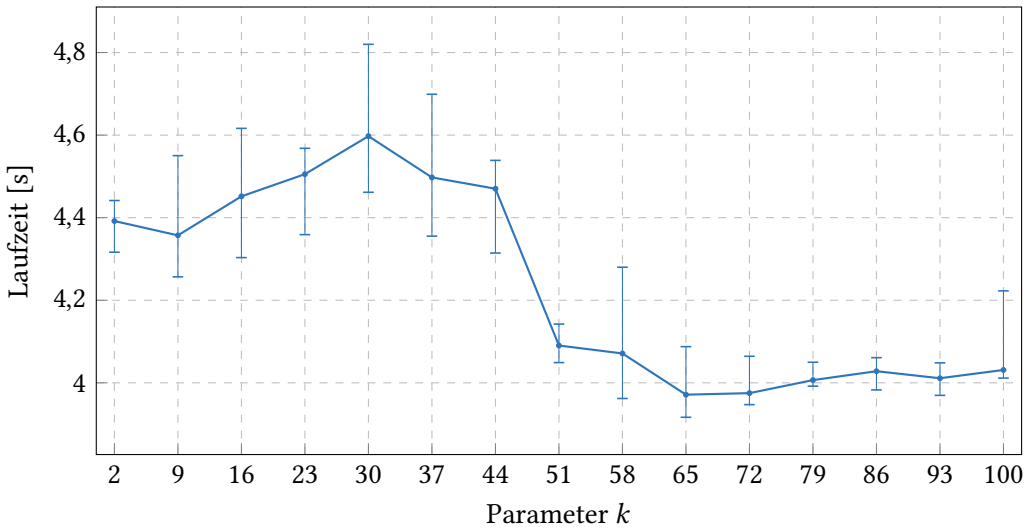

Abbildung 5.3.: Ausführungszeiten für verschiedene Werte des Parameters k

verschiedenen Werte bei zehn Durchläufen zwar zu messbaren Laufzeitdifferenzen führen, deren relative Abweichungen jedoch recht gering sind. Eine Analyse für den konkreten Datensatz ergab, dass ab dem Wert $k = 51$ das Attribut Alter zu „$*$“ generalisiert wird, was *ARX* vermutlich das Erzeugen der Äquivalenzklassen erleichtert. Insgesamt impliziert die ohnehin niedrige Ausführungszeit keine Limitierungen bezüglich der Wahl des Parameters k.

5.2. Präzision im Kontext von Differential Privacy

Für das im interaktiven Szenario eingesetzte Konzept der DP sollte der Einfluss des Parameters ε auf die Abweichung des von PRIORI zurückgegebenen Ergebnisses vom exakten Ergebnis ermittelt werden. Hierzu wurde ein Datensatz mit 10 000 Einträgen erzeugt (`interactive.csv`) und sodann mit der SQL-Anfrage `SELECT COUNT(*) FROM PRIORI` bei $\varepsilon \in \{10^{0{,}01x} \mid x \in [-200, 100]\}$ und festem $\delta = 10^{-12}$ fünfmal ausgewertet. Schaubild 5.4 zeigt, dass die Streuung um das tatsächliche Ergebnis (10 000) gemäß der Laplace-Verteilung mit steigendem ε abnimmt (siehe Abschnitt 3.2.5.3). Ferner scheinen die in der Literatur empfohlenen Werte zwischen 0,01 und $\ln 3$ den sinnvollen Bereich korrekt einzugrenzen.

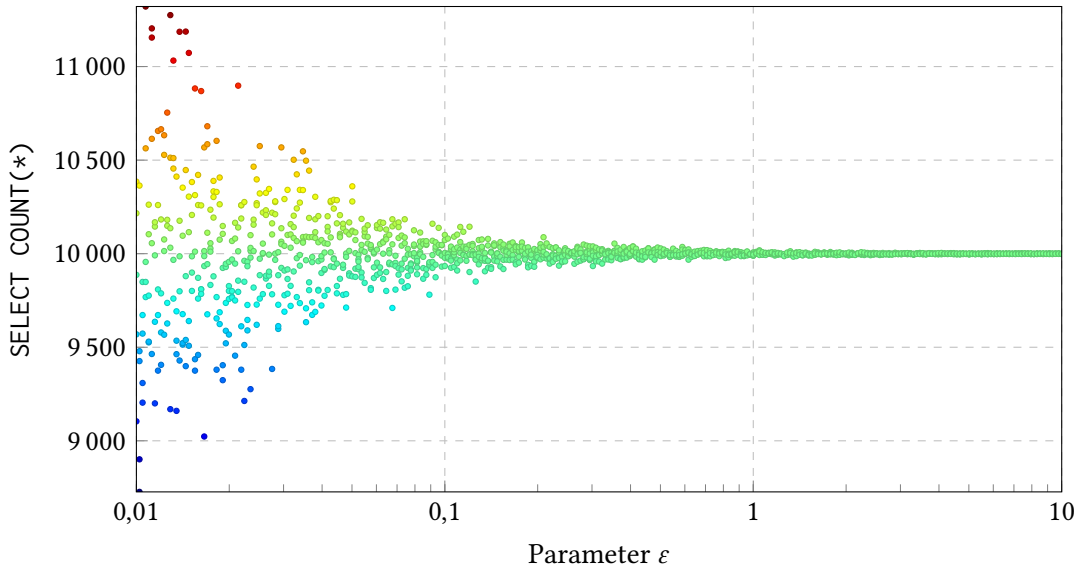

Abbildung 5.4.: Einfluss des Parameters ε auf das Ergebnis

5.3. Diskussion

Der in der vorliegenden Arbeit bei der prototypischen Forschungsschnittstelle gewählte Ansatz ist mit gewissen Möglichkeiten und Einschränkungen verbunden, die nachfolgend aufgezeigt

werden sollen.

Zuvorderst setzt PRIORI umfangreiches Wissen für eine Freigabe von Datensätzen zu [medizinischen] Forschungszwecken voraus. Zum einen müssen Nutzer PRIORI entweder direkt über das REST-API oder über das Swagger UI ansprechen, da bislang keine intuitive Benutzerschnittstelle zur Verfügung steht. Zum anderen muss der Medizininformatiker die unterschiedlichen PETs im Detail verstanden haben, um im Rahmen einer Güterabwägung den Zielkonflikt zwischen Datenschutz und Forschungsdatennutzung durch die Wahl geeigneter Parameter bestmöglich zu lösen. Sicherlich könnte PRIORI den Medizininformatiker bei dieser Entscheidung noch umfassender unterstützen, auch wenn mit der Möglichkeit einer gewichteten Risikobewertung bereits ein erster Schritt in diese Richtung unternommen wurde. Von einer vollständig automatisierten Wahl der Parameter durch PRIORI ist allerdings wegen der hieraus resultierenden Gefahren in Anbetracht der besonderen Schutzwürdigkeit von Gesundheitsdaten abzuraten.

Dieser Ansatz stellt insofern eine Einschränkung dar, dass nur wenige Einrichtungen solches Fachpersonal beschäftigen. Ein Medizininformatiker mit den nötigen Kenntnissen dürfte noch am ehesten an Universitätskliniken oder Forschungsdatenzentren anzutreffen sein. Demgegenüber verfügen derartige Einrichtungen über entsprechende Gesundheitsdaten, die im Zuge einer Datenspende durch die Forschung analysiert werden können.

Gleichzeitig zeigen die vorangegangenen Ausführungen, dass wegen der gegebenen Performanz selbst größere Datenmengen PRIORI nicht vor unmögliche Herausforderungen stellen. Mit k-Anonymity, den verschiedenen Formen von ℓ-Diversity und t-Closeness sowie der DP werden die bedeutendsten PETs unterstützt, weitere Funktionen können wegen des modularen Aufbaus von PRIORI mit überschaubarem Aufwand ergänzt werden. Im Gegensatz zu verwandten Arbeiten sind mit Authentifizierung und Zugriffskontrolle bereits wichtige Datenschutzmechanismen implementiert. Des Weiteren kann ein Datensatz nach einmaligem Hochladen im Rahmen mehrerer Projekte mit individuellen Parametern für die Forschung nutzbar gemacht werden. Durch die Verwendung des REST-Architekturstils ist schon jetzt eine Trennung der Datenhaltung von der Benutzerschnittstelle und damit das fundamentale Prinzip der *Separation of Concerns* gewährleistet.

6. Fazit und Ausblick

Im folgenden Kapitel werden die in dieser Arbeit gewonnenen Erkenntnisse zusammengefasst und ein Ausblick auf zukünftige Arbeiten gegeben.

6.1. Fazit

Die Digitalisierung macht auch vor dem Gesundheitswesen nicht halt. Aus diesem Grund war es das Ziel, zu Beginn dieser Arbeit einen Überblick über den aktuellen Stand von E-Health in Deutschland zu bieten. Hierzu wurde auf die Telematikinfrastruktur eingegangen, welche alle Beteiligten im Gesundheitswesen miteinander vernetzt und somit das Fundament für die in der Einführung befindliche ePA bildet. Daneben wurden die Inhalte und die Zugriffskonzeption der ePA analysiert sowie die datenschutzrechtlichen Grundlagen für die Forschung mit Gesundheitsdaten beleuchtet.

Anschließend wurden existierende PETs ausführlich erläutert. Anhand der erwähnten Re-Identifizierungen wurde ersichtlich, dass bei Gesundheitsdaten eine Pseudonymisierung oder das Entfernen eindeutiger Identifikationsmerkmale regelmäßig nicht ausreicht, um einen angemessenen Schutz der Privatsphäre zu garantieren. Vor diesem Hintergrund wurden in der Literatur Anonymitätsmaße wie k-Anonymity, ℓ-Diversity, t-Closeness, δ-Presence und DP entwickelt, deren jeweilige Stärken und Schwächen aufgezeigt wurden. Weiterhin wurden die zugehörigen Anonymisierungsverfahren behandelt. Als Alternative zur Anonymisierung wurde die homomorphe Verschlüsselung vorgestellt, die nicht auf Einträge aus relationalen Datenbanken beschränkt ist, sondern beliebige Berechnungen auf verschlüsselten Daten ermöglicht. Um den Umfang dieser Arbeit nicht zu sprengen, wurde sie im Folgenden nicht weiter betrachtet.

Mittels PRIORI, einer im Rahmen der vorliegenden Arbeit entwickelten prototypischen Forschungsschnittstelle auf Basis des REST-Architekturstils, wurden die bedeutendsten PETs hinsichtlich ihrer Eignung für medizinische Daten untersucht. Die Evaluierung zeigte, dass PRIORI mithilfe von PETs, die bereits als Open-Source-Software implementiert wurden, eine Anonymisierung und statistische Auswertung von Datensätzen in angemessener Laufzeit erlaubt. Abschließend wurde die Laufzeit in Abhängigkeit des Parameters k von k-Anonymity

und der Einfluss des Parameters ε von DP auf die Präzision der zurückgegebenen Ergebnisse analysiert.

6.2. Ausblick

Bei PRIORI handelt es sich ungeachtet der gegebenen Performanz um einen Prototyp, für den es Erweiterungs- und Verbesserungspotenziale gibt. Zum einen könnten weitere Transformationsarten bei der Erzeugung von Generalisierungshierarchien unterstützt werden, denn bislang bietet PRIORI nur das schrittweise Entfernen von Informationen (sog. *Redaction*) oder die Abbildung einzelner Werte auf Intervalle an. Beispielsweise könnte eine Transformation, welche die Generierung von Hierarchien für *Timestamps* ermöglicht, hilfreich sein. Für umfangreiche Datensätze könnte außerdem eine separate Übermittlung der Generalisierungshierarchien als CSV-Datei effizienter sein.

Zum anderen besteht noch nicht die Möglichkeit, den Parameter ε von DP als *Privacy Budget* zu verwenden. Somit sind aktuell im interaktiven Szenario beliebig viele Anfragen möglich. Darüber hinaus könnte die Unterstützung der Nutzer bei der Parameterwahl verbessert werden. Ferner kann eine graphische Benutzerschnittstelle für PRIORI angesichts der gewählten REST-Architektur unkompliziert ergänzt werden. Diese Maßnahme würde die Zielgruppe von PRIORI erweitern, da ein wesentlich geringeres technisches Verständnis vonnöten wäre.

Die Frage nach der Anwendbarkeit der PETs auf Gesundheitsdaten sollte grundsätzlich durch interdisziplinäre Forschung vertieft werden, da die vorliegende Arbeit lediglich mathematisch-naturwissenschaftliche und ingenieurwissenschaftliche Aspekte thematisiert. In diesem Zusammenhang sollten Lösungsmöglichkeiten für die Tatsache erarbeitet werden, dass die Daten oftmals nur in textbasierter und nicht in strukturierter Form vorliegen und aus diesem Grund bisher nur mit großem Aufwand ausgewertet werden können. Zudem sollten alternative Herangehensweisen wie *Federated Learning*,[1] das eine dezentrale Analyse von Datensätzen mithilfe von maschinellem Lernen erlaubt, und neuartige PETs Gegenstand weiterer Untersuchungen sein.

1 `https://federated.withgoogle.com` (besucht am 08. 10. 2020)

Literatur

[Aca+18] Abbas Acar et al. “A Survey on Homomorphic Encryption Schemes.” In: *ACM Computing Surveys* 51.4 (Sept. 2018), pp. 1–35. DOI: 10.1145/3214303.

[AH16] Volker P. Andelfinger und Till Hänisch, Hrsg. *eHealth*. Springer Fachmedien Wiesbaden, 2016. DOI: 10.1007/978-3-658-12239-3.

[Ärz20] Ärzte Zeitung online. *gematik kündigt medizinische Anwendungen im Juli an*. 25. Mai 2020. URL: https://www.aerztezeitung.de/Wirtschaft/gematik-kuendigt-medizinische-Anwendungen-im-Juli-an-409747.htm (besucht am 29. 05. 2020).

[BEE18] Christoph Bauer, Frank Eickmeier und Michael Eckard. *E-Health: Datenschutz und Datensicherheit*. Springer Fachmedien Wiesbaden, 2018. DOI: 10.1007/978-3-658-15091-4.

[BfDI20] Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit. *Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche*. 29. Juni 2020. URL: https://www.bfdi.bund.de/DE/Infothek/Transparenz/Konsultationsverfahren/01_Konsulation-Anonymisierung-TK/Positionspapier-Anonymisierung.pdf.

[BK20] Rebecca Beerheide und Heike E. Krüger-Brand. „Patientendaten-Schutz-Gesetz: E-Rezept und E-Akte im Fokus“. In: *Dtsch Arztebl International* 117.15 (2020), A-756–A-757. URL: https://www.aerzteblatt.de/int/article.asp?id=213474.

[BLD97] Der Bayerische Landesbeauftragte für den Datenschutz. *Arbeitspapier Datenschutzfreundliche Technologien*. 17. Nov. 1997. URL: https://www.datenschutz-bayern.de/technik/grundsatz/apdsft.htm (besucht am 15. 07. 2020).

[BS08] Justin Brickell and Vitaly Shmatikov. “The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing.” In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’08*. Las Vegas, USA: ACM Press, 2008, pp. 70–78. DOI: 10.1145/1401890.1401904.

[BT20] Deutscher Bundestag. *Drucksache 19/18793: Entwurf eines Gesetzes zum Schutz elektronischer Patientendaten in der Telematikinfrastruktur*. 27. Apr. 2020. URL: https://dip21.bundestag.de/dip21/btd/19/187/1918793.pdf (besucht am 02. 06. 2020).

[Cic+19] Eleonora Ciceri et al. "PAPAYA: A platform for privacy preserving data analytics." In: *ERCIM News, Special theme: Digital health* 118 (July 2019). URL: https://ercim-news.ercim.eu/en118/special/papaya-a-platform-for-privacy-preserving-data-analytics.

[CK12] Jianneng Cao and Panagiotis Karras. "Publishing Microdata with a Robust Privacy Guarantee." In: *Proceedings of the VLDB Endowment* 5.11 (July 2012), pp. 1388–1399. DOI: 10.14778/2350229.2350255.

[CT13] Chris Clifton and Tamir Tassa. "On Syntactic Anonymity and Differential Privacy." In: *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*. Brisbane, Australia: IEEE, Apr. 2013, pp. 88–93. DOI: 10.1109/ICDEW.2013.6547433.

[Dal77] Tore Dalenius. "Towards a methodology for statistical disclosure control." In: *Statistik Tidskrift* 15 (1977), pp. 429–444.

[DE13] Fida K. Dankar and Khaled El Emam. "Practicing Differential Privacy in Health Care: A Review." In: *Transactions on Data Privacy* 6.1 (Apr. 2013), pp. 35–67. URL: https://www.tdp.cat/issues11/tdp.a129a13.pdf.

[Deu18] Deutsches Ärzteblatt. *Die Gesundheitsakte ist eine Übergangslösung*. 30. Juli 2018. URL: https://www.aerzteblatt.de/nachrichten/96793/Die-Gesundheitsakte-ist-eine-Uebergangsloesung (besucht am 06. 06. 2020).

[Dow+17] Nathan Dowlin et al. "Manual for Using Homomorphic Encryption for Bioinformatics." In: *Proceedings of the IEEE* (2017), pp. 1–16. DOI: 10.1109/jproc.2016.2622218.

[DPT17] Differential Privacy Team, Apple. *Learning with Privacy at Scale*. Dec. 2017. URL: https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

[DR13] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy." In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4 (2013), pp. 211–407. DOI: 10.1561/0400000042.

[DR20] Christian Dierks und Alexander Roßnagel. *Sekundärnutzung von Sozial- und Gesundheitsdaten*. Medizinisch Wissenschaftliche Verlagsgesellschaft, Feb. 2020. DOI: 10.32745/9783954665181.

[DSH19] Richard Dosselmann, Mehdi Sadeqi, and Howard J. Hamilton. *A Tutorial on Computing t-Closeness*. Nov. 25, 2019. arXiv: 1911.11212 [cs.CR].

[DSK20] Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder. *Standard-Datenschutzmodell*. 17. Apr. 2020. URL: https://www.bfdi.bund.de/SharedDocs/Publikationen/Sachthemen/Standard-Datenschutzmodell.pdf.

[Dwo+06] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis." In: *Theory of Cryptography*. Springer Berlin Heidelberg, 2006, pp. 265–284. DOI: 10.1007/11681878_14.

[Dwo06] Cynthia Dwork. "Differential Privacy." In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Springer Berlin Heidelberg, 2006, pp. 1–12. DOI: 10.1007/11787006_1.

[Dwo08] Cynthia Dwork. "Differential Privacy: A Survey of Results." In: *Lecture Notes in Computer Science*. Ed. by Manindra Agrawal et al. Springer Berlin Heidelberg, 2008, pp. 1–19. DOI: 10.1007/978-3-540-79228-4_1.

[Ema+09] Khaled El Emam et al. "A Globally Optimal *k*-Anonymity Method for the De-Identification of Health Data." In: *Journal of the American Medical Informatics Association* 16.5 (Sept. 2009), pp. 670–682. DOI: 10.1197/jamia.m3144.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*. Scottsdale, USA: ACM Press, 2014, pp. 1054–1067. DOI: 10.1145/2660267.2660348.

[EY18] Ernst & Young. *When the human body is the biggest data platform, who will capture value?* 2018. URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/digital/ey-when-the-human-body-is-the-biggest-data-platform-who-will-capture-value.pdf.

[FG07] Caroline Fontaine and Fabien Galand. "A Survey of Homomorphic Encryption for Nonspecialists." In: *EURASIP Journal on Information Security* 2007 (2007), pp. 1–10. DOI: 10.1155/2007/13801.

[Fie00] Roy Thomas Fielding. "Architectural Styles and the Design of Network-Based Software Architectures." PhD thesis. University of California, Irvine, 2000. URL: https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.

[FK16] Florian Fischer und Alexander Krämer, Hrsg. *eHealth in Deutschland*. Springer Berlin Heidelberg, 2016. DOI: 10.1007/978-3-662-49504-9.

[FP01] Curt D. Furberg and Bertram Pitt. "Withdrawal of cerivastatin from the world market." In: *Current Controlled Trials in Cardiovascular Medicine* 2.5 (2001), p. 205. DOI: 10.1186/cvm-2-5-205.

[Fun+10] Benjamin C. M. Fung et al. "Privacy-Preserving Data Publishing: A Survey of Recent Developments." In: *ACM Computing Surveys* 42.4 (June 2010), pp. 1–53. DOI: 10.1145/1749603.1749605.

[gem19a] gematik GmbH. *Konzept Architektur der TI-Plattform (gemKPT_Arch_TIP)*. Version 2.9.0. 2. Okt. 2019. URL: https://www.vesta-gematik.de/standard/formhandler/324/gemKPT_Arch_TIP_V2_9_0.pdf.

[gem19b] gematik GmbH. *Spezifikation Schlüsselgenerierungsdienst ePA (gemSpec_SGD_ePA)*. Version 1.2.0. 9. Okt. 2019. URL: https://www.vesta-gematik.de/standard/formhandler/324/gemSpec_SGD_ePA_V1_2_0.pdf.

[gem19c] gematik GmbH. *Systemspezifisches Konzept ePA (gemSysL_ePA)*. Version 1.3.0. 2. Okt. 2019. URL: https://www.vesta-gematik.de/standard/formhandler/324/gemSysL_ePA_V1_3_0.pdf.

[gem19d] gematik GmbH. *Whitepaper: Datenschutz und Informationssicherheit in der Telematikinfrastruktur*. Nov. 2019.

[Gen09] Craig Gentry. "Fully Homomorphic Encryption Using Ideal Lattices." In: *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing - STOC '09*. Bethesda, USA: ACM Press, 2009, pp. 169–178. DOI: 10.1145/1536414.1536440.

[Gil+16] Ran Gilad-Bachrach et al. "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy." In: *Proceedings of the 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, USA: PMLR, June 2016, pp. 201–210. URL: http://proceedings.mlr.press/v48/gilad-bachrach16.pdf.

[GL15] Aris Gkoulalas-Divanis and Grigorios Loukides, eds. *Medical Data Privacy Handbook*. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-23633-9.

[GLS14] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun. "Publishing data from electronic health records while preserving privacy: A survey of algorithms." In: *Journal of Biomedical Informatics* 50 (Aug. 2014), pp. 4–19. DOI: 10.1016/j.jbi.2014.06.002.

[JS20] Alexandra Jorzig und Frank Sarangi. *Digitalisierung im Gesundheitswesen*. Springer Berlin Heidelberg, 2020. DOI: 10.1007/978-3-662-58306-7.

[Koc+13] Ovunc Kocabas et al. "Assessment of Cloud-based Health Monitoring using Homomorphic Encryption." In: *2013 IEEE 31st International Conference on Computer Design (ICCD)*. Asheville, USA: IEEE, Oct. 2013, pp. 443–446. DOI: 10.1109/iccd.2013.6657078.

[Krü15] Heike E. Krüger-Brand. „E-Health-Gesetz: Wichtige Etappe erreicht". In: *Dtsch Arztebl International* 112.50 (2015), A-2124–A-2126. URL: https://www.aerzteblatt.de/int/article.asp?id=173307.

[Krü20] Heike E. Krüger-Brand. „Kommunikation: E-Arztbrief im Feldtest". In: *Dtsch Arztebl International* 117.19 (2020), A-988–A-989. URL: https://www.aerzteblatt.de/int/article.asp?id=213896.

[Küh19] Jürgen Kühling. „Datenschutz im Gesundheitswesen". In: *Medizinrecht* 37.8 (Aug. 2019), S. 611–622. DOI: 10.1007/s00350-019-5291-y.

[LBÜ15] Martin Lablans, Andreas Borg, and Frank Ückert. "A RESTful interface to pseudonymization services in modern web applications." In: *BMC Medical Informatics and Decision Making* 15.1 (Feb. 2015). DOI: 10.1186/s12911-014-0123-5.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity." In: *2007 IEEE 23rd International Conference on Data Engineering*. Istanbul, Turkey: IEEE, Apr. 2007, pp. 106–115. DOI: 10.1109/icde.2007.367856.

[Mac+07] Ashwin Machanavajjhala et al. "ℓ-diversity: Privacy Beyond k-Anonymity." In: *ACM Transactions on Knowledge Discovery from Data* 1.1 (Mar. 2007), p. 3. DOI: 10.1145/1217299.1217302.

[Mas11] Mark Masse. *REST API Design Rulebook*. O'Reilly Media, Inc, USA, Nov. 11, 2011. ISBN: 1449310508.

[May20] Thorsten Maybaum. „E-Health: Praxisausweise können wieder bestellt werden“. In: *Dtsch Arztebl International* 117.5 (2020), A–182. URL: https://www.aerzteblatt.de/int/article.asp?id=212266.

[MHN15] Jörn Müller-Quade, Matthias Huber und Tobias Nilges. „Daten verschlüsselt speichern und verarbeiten in der Cloud“. In: *Datenschutz und Datensicherheit - DuD* 39.8 (Aug. 2015), S. 531–535. DOI: 10.1007/s11623-015-0465-x.

[ML17] Stefan Müller-Mielitz und Thomas Lux, Hrsg. *E-Health-Ökonomie*. Springer Fachmedien Wiesbaden, 2017. DOI: 10.1007/978-3-658-10788-8.

[MPE17] David Matusiewicz, Christian Pittelkau und Arno Elmer, Hrsg. *Die Digitale Transformation im Gesundheitswesen*. Medizinisch Wissenschaftliche Verlagsgesellschaft, Sep. 2017. DOI: 10.32745/9783954663576.

[MT07] Frank McSherry and Kunal Talwar. “Mechanism Design via Differential Privacy.” In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07)*. Providence, USA: IEEE, Oct. 2007, pp. 94–103. DOI: 10.1109/focs.2007.66.

[NAC07] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. “Hiding the presence of individuals from shared databases.” In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD ’07*. Beijing, China: ACM Press, 2007, pp. 665–676. DOI: 10.1145/1247480.1247554.

[NC10] Mehmet Ercan Nergiz and Chris Clifton. “δ-Presence without Complete World Knowledge.” In: *IEEE Transactions on Knowledge and Data Engineering* 22.6 (June 2010), pp. 868–883. DOI: 10.1109/tkde.2009.125.

[NS08] Arvind Narayanan and Vitaly Shmatikov. “Robust De-anonymization of Large Sparse Datasets.” In: *2008 IEEE Symposium on Security and Privacy*. Oakland, USA: IEEE, May 2008, pp. 111–125. DOI: 10.1109/sp.2008.33.

[NW18] Kobbi Nissim and Alexandra Wood. “Is privacy privacy?” In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (Aug. 2018). DOI: 10.1098/rsta.2017.0358.

[PAP19] PAPAYA project. *Requirements Specification*. Version 1.1. Apr. 29, 2019. URL: https://www.papaya-project.eu/sites/default/files/papaya/public/content-files/deliverables/PAPAYA_D2_2_Requirements_Specification.pdf.

[PR04] Klaus Pommerening and Michael Reng. "Secondary Use of the EHR via Pseudonymisation." In: *Studies in Health Technology and Informatics* 103 (May 2004), pp. 441–446. DOI: 10.3233/978-1-60750-946-2-441.

[Pra+20] Fabian Prasser et al. "Flexible data anonymization using ARX—Current status and challenges ahead." In: *Software: Practice and Experience* 50.7 (Feb. 2020), pp. 1277–1304. DOI: 10.1002/spe.2812.

[PS17] Ronald Petrlic und Christoph Sorge. *Datenschutz*. Springer Fachmedien Wiesbaden, 2017. DOI: 10.1007/978-3-658-16839-1.

[RAD78] Ronald L. Rivest, Len Adleman, and Michael L. Dertouzos. "On Data Banks and Privacy Homomorphism." In: *Foundations on Secure Computation.* Academia Press, 1978, pp. 169–179.

[RHM19] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. "Estimating the success of re-identifications in incomplete datasets using generative models." In: *Nature Communications* 10.1 (July 2019). DOI: 10.1038/s41467-019-10933-3.

[Sch16] Uwe Klaus Schneider. *Einrichtungsübergreifende elektronische Patientenakten*. Springer Fachmedien Wiesbaden, 2016. DOI: 10.1007/978-3-658-11597-5.

[SS98] Pierangela Samarati and Latanya Sweeney. *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression*. 1998. URL: https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf.

[SW19] Rolf Schwartmann und Steffen Weiß. *Anforderungen an den datenschutzkonformen Einsatz von Pseudonymisierungslösungen*. Version 1.01. Gesellschaft für Datenschutz und Datensicherheit e.V. 2019. URL: https://www.gdd.de/downloads/anforderungen-an-datenschutzkonforme-pseudonymisierung.

[Swe02] Latanya Sweeney. "*k*-Anonymity: A Model for Protecting Privacy." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (Oct. 2002), pp. 557–570. DOI: 10.1142/s0218488502001648.

[Viz+19] Anamaria Vizitiu et al. "Towards Privacy-Preserving Deep Learning based Medical Imaging Applications." In: *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. Istanbul, Turkey: IEEE, June 2019, pp. 1–6. DOI: 10.1109/memea.2019.8802193.

[Wey20] Jens Weyd. „Digitalisierung in der Gesetzlichen Krankenversicherung". In: *Medizinrecht* 38.3 (März 2020), S. 183–192. DOI: 10.1007/s00350-020-5480-8.

[Wil+20] Royce J Wilson et al. “Differentially Private SQL with Bounded User Contribution.” In: *Proceedings on Privacy Enhancing Technologies* 2020.2 (Apr. 2020), pp. 230–250. DOI: 10.2478/popets-2020-0025.

[Zig+20] Athanasios Zigomitros et al. “A Survey on Privacy Properties for Data Publishing of Relational Data.” In: *IEEE Access* 8 (2020), pp. 51071–51099. DOI: 10.1109/access.2020.2980235.

Anhang

A. POST /api/v1/login

```
{
    "email_address": "user@example.com",
    "password": "123456"
}
```

JSON-Objekt 1.: Beispielhafte Anfrage

```
{
    "access_token": "JWT",
    "refresh_token": "JWT"
}
```

JSON-Objekt 2.: Beispielhafte Antwort

B. POST /api/v1/users

```
{
    "email_address": "jane@example.com",
    "institution": "Fraunhofer IOSB",
    "name": "Jane Doe",
    "password": "123456",
    "privileged_status": false
}
```

JSON-Objekt 3.: Beispielhafte Anfrage

```
{
    "email_address": "jane@example.com",
    "institution": "Fraunhofer IOSB",
    "name": "Jane Doe",
    "privileged_status": false,
    "user_id": 3
}
```


JSON-Objekt 4.: Beispielhafte Antwort

C. POST /api/v1/datasets


```
{
    "name": "Heart Diseases",
    "description": "Heart diseases among German residents"
}
```


JSON-Objekt 5.: Beispielhafte Anfrage


```
{
    "created_at": "2020-09-30T19:06:00.909045",
    "creator": {
        "email_address": "john@example.com",
        "institution": "Fraunhofer IOSB",
        "name": "John Doe",
        "privileged_status": false,
        "user_id": 2
    },
    "dataset_id": 1,
    "description": "Heart diseases among German residents",
    "name": "Heart Diseases"
}
```


JSON-Objekt 6.: Beispielhafte Antwort

D. POST /api/v1/datasets/<*dataset_id*>/create-generalization-hierarchy

```
{
    "transformation_type": "redaction",
    "attribute_name": "zip_code",
    "redaction_order": "right_to_left"
}
```

JSON-Objekt 7.: Beispielhafte Anfrage (`transformation_type = redaction`)

```
{
    "hierarchy": [
        [
            "81667",
            "8166*",
            "816**",
            "81***",
            "8****",
            "*****"
        ],
        [
            "81668",
            "8166*",
            "816**",
            "81***",
            "8****",
            "*****"
        ]
    ]
}
```

JSON-Objekt 8.: Beispielhafte Antwort (`transformation_type = redaction`)

```
{
    "attribute_name": "age",
    "transformation_type": "interval",
    "intervals": [
        {
            "label": "young",
            "min": 0,
            "max": 18
        },
        {
            "label": "adult",
            "min": 18,
            "max": 70
        },
        {
            "label": "elderly",
            "min": 70,
            "max": 100
        }
    ]
}
```

JSON-Objekt 9.: Beispielhafte Anfrage (`transformation_type = interval`)

```
{
    "hierarchy": [
        [
            "13",
            "young",
            "*"
        ],
        [
            "28",
            "adult",
            "*"
        ],
        [
            "71",
            "elderly",
            "*"
        ]
    ]
}
```

JSON-Objekt 10.: Beispielhafte Antwort (`transformation_type = interval`)

E. POST /api/v1/datasets/<*dataset_id*>/analyze-risk

```
{
    "attributes": [
        {
            "attribute_name": "zip_code",
            "attribute_type": "quasi-identifying",
            "attribute_hierarchy": [
                [
                    "476",
                    "47*",
                    "4**",
                    "***"
                ],
                ...
            ]
        },
        {
            "attribute_name": "age",
            "attribute_type": "quasi-identifying",
            "attribute_hierarchy": [
                [
                    "30",
                    "adult",
                    "*"
                ],
                ...
            ]
        },
        {
            "attribute_name": "disease",
            "attribute_type": "sensitive"
        }
    ],
    "privacy_models": [
        {
            "model_name": "k-anonymity",
            "model_params": {
                "k": 5
            }
        },
        {
            "model_name": "l-diversity_distinct",
            "attribute_name": "disease",
            "model_params": {
                "l": 2
            }
        }
    ],
    "suppression_limit": 0.2
}
```

JSON-Objekt 11.: Beispielhafte Anfrage

```
{
    "attribute_generalization": [
        {
            "attribute_name": "zip_code",
            "generalization_level": 3
        },
        {
            "attribute_name": "age",
            "generalization_level": 3
        }
    ],
    "risk_profile": {
        "attacker_success_rate": {
            "journalist_attacker_success_rate": 0.1,
            "marketer_attacker_success_rate": 0.1,
            "prosecutor_attacker_success_rate": 0.1
        },
        "distribution_of_risk": [
            {
                "interval": "[50,100]",
                "records_with_maximal_risk_within_interval": 1,
                "records_with_risk_within_interval": 0
            },
            {
                "interval": "[33.4,50)",
                "records_with_maximal_risk_within_interval": 1,
                "records_with_risk_within_interval": 0
            },
            ...
        ],
        "re_identification_risk": {
            "average_prosecutor_risk": 0.1,
            "estimated_journalist_risk": 0.1,
            "estimated_marketer_risk": 0.1,
            "estimated_prosecutor_risk": 0.1,
            "highest_journalist_risk": 0.1,
            "highest_prosecutor_risk": 0.1,
            "lowest_risk": 0.1,
            "population_uniques": 0,
            "records_affected_by_highest_journalist_risk": 1,
            "records_affected_by_highest_prosecutor_risk": 1,
            "records_affected_by_lowest_risk": 1,
            "sample_uniques": 0
        }
    }
}
```


JSON-Objekt 12.: Beispielhafte Antwort

F. POST /api/v1/datasets/<*dataset_id*>/calculate-weighted-risk


```
{
    "risk": 0.19393939393939394
}
```


JSON-Objekt 13.: Beispielhafte Antwort für die Anfrage aus Abschnitt E

G. POST /api/v1/projects


```
{
    "name": "Heart Diseases",
    "description": "Heart diseases among German residents",
    "dataset_id": 1,
    "researcher_id": 3,
    "config": {
        "attributes": [
            {
                "attribute_name": "age",
                "attribute_type": "quasi-identifying",
                "metadata": {
                    "datatype": "int",
                    "min": 0,
                    "max": 100
                }
            },
            {
                "attribute_name": "gender",
                "attribute_type": "quasi-identifying",
                "metadata": {
                    "datatype": "str",
                    "cardinality": 2
                }
            }
        ],
        "privacy_models": [
            {
                "model_name": "epsilon-delta-differential_privacy",
                "model_params": {
                    "epsilon": 2,
                    "delta": 10E-16
                }
            }
        ]
    }
}
```


JSON-Objekt 14.: Beispielhafte Anfrage

```
{
    "attributes": [
        {
            "attribute_name": "age",
            "metadata": {
                "datatype": "int",
                "min": 0,
                "max": 100
            }
        },
        {
            "attribute_name": "gender",
            "metadata": {
                "cardinality": 2,
                "datatype": "str"
            }
        }
    ],
    "created_at": "2020-10-01T10:18:03.374430",
    "creator": {
        "email_address": "john@example.com",
        "institution": "Fraunhofer IOSB",
        "name": "John Doe",
        "privileged_status": true,
        "user_id": 2
    },
    "dataset_id": 1,
    "description": "Heart diseases among German residents",
    "name": "Heart Diseases",
    "project_id": 4,
    "researcher": {
        "email_address": "jane@example.com",
        "institution": "Fraunhofer IOSB",
        "name": "Jane Doe",
        "privileged_status": false,
        "user_id": 3
    }
}
```


JSON-Objekt 15.: Beispielhafte Antwort

H. POST /api/v1/projects/<*project_id*>/analyze

```
{
    "query": "SELECT married, AVG(age) FROM PRIORI GROUP BY married"
}
```

JSON-Objekt 16.: Beispielhafte Anfrage

```
{
    "query_result": {
        "Col1": {
            "0": 39.04143174415034,
            "1": 46.140899295898265
        },
        "married": {
            "0": false,
            "1": true
        }
    }
}
```

JSON-Objekt 17.: Beispielhafte Antwort