Quantifizierung und Schutz der Privatsphäre in der (Epi-)genetik[1]

Pascal Berrang[2]

Abstract: Die stetig sinkenden Kosten für molekulares Profiling haben der Biomedizin zahlreiche neue Arten von Daten geliefert und den Durchbruch für eine präzisere und personalisierte Medizin ermöglicht. Die Veröffentlichung dieser inhärent hochsensiblen Daten stellt jedoch eine neue Bedrohung für unsere Privatsphäre dar. Während die IT-Sicherheitsforschung sich bisher hauptsächlich auf die Auswirkungen *genetischer* Daten auf die Privatsphäre konzentriert hat, wurden die vielfältigen Risiken durch andere Arten biomedizinischer Daten größtenteils außer Acht gelassen.

Wir stellen Verfahren zur Messung und Abwehr solcher Privatsphärerisiken vor. Neben dem Genom konzentrieren wir uns auf zwei der wichtigsten gesundheitsrelevanten epigenetischen Elemente: microRNAs und DNA-Methylierung. Wir quantifizieren die Privatsphäre für mehrere realistische Angriffsszenarien. Unsere Resultate bekräftigen, dass die Privatsphärerisiken solcher Daten ernst zu nehmen sind. Zudem präsentieren und evaluieren wir Lösungen zum Schutz der Privatsphäre. Sie reichen von der Anwendung von Differential Privacy bis zu kryptographischen Protokollen.

1 Einführung

Seit der ersten vollständigen Genomsequenzierung im Jahr 2001 sind die Kosten für molekulares Profiling stetig gesunken und haben somit den Durchbruch für eine präzisere und personalisierte Medizin ermöglicht. Während dieser Durchbruch durch den enormen Zuwachs an verfügbaren biomedizinischen Daten begleitet und bestärkt wird, zeigt dies auch gleichzeitig die Schattenseite der Entwicklung: neue Privatsphärerisiken im Gesundheitsbereich. Unser Genom ist in dieser Hinsicht besonders betroffen, da es uns nicht nur eindeutig identifiziert, sondern sich auch während des gesamten Lebens kaum verändert. Außerdem lassen sich aufgrund der Vererbung des Genoms sogar Rückschlüsse auf Verwandte ziehen [Hu13]. Diese Besonderheiten erklären vermutlich, wieso sich die IT-Sicherheitsgemeinde bisher hauptsächlich auf die Implikationen *genetischer* Daten fokussiert hat. Verschiedene Angriffsvektoren und Schutzmaßnahmen in diesem Bereich wurden bereits 2014 weitreichend untersucht und kategorisiert [EN14].

Das Genom ist jedoch nicht der einzige Bestandteil des menschlichen Körpers, welcher für unsere Gesundheit von Bedeutung ist. Umgebungseinflüsse wie Umweltverschmutzung, unsere Ernährung und unser Lebensstil sind oft ausschlaggebend für die Entwicklung der meisten Erkrankungen. Multi-omik Ansätze wie Epigenetik, Transkriptomik und Proteomik versprechen genau diese Lücke zwischen dem Genom und unserem Gesundheitszustand zu schließen. Während unser Genom kodiert wie sich Zellen potentiell verhalten können, geben das Epigenom und Transkriptom Aufschluss darüber wie sich eine

[1] Englischer Titel der Dissertation: “Quantifying and Mitigating Privacy Risks in Biomedical Data”

[2] CISPA, Universität des Saarlandes, contact@paberr.net

einzelne Zelle zum aktuellen Zeitpunkt wirklich verhält. In einer Computer Analogie zusammengefasst: wenn das Genom unsere Hardware ist, dann entspricht das Epigenom unserer Software [Cl10]. Genau wie unser Gesundheitszustand variieren epigenetische Daten daher über die Zeit und werden von der Umgebung beeinflusst.

Obwohl die Epigenetik in der Biomedizin ständig an Bedeutung dazu gewinnt, wurden damit verbundene Privatsphärerisiken bisher größtenteils außer Acht gelassen. Mit dem wachsenden Verständnis für epigenetische Daten wird allerdings klar, welche Sensibilität die darin enthaltenen Informationen haben. Es wurden nicht nur Verbindungen zu einer Reihe schwerer Erkrankungen (wie Krebs, Diabetes oder Alzheimer [Wo07, JB07, QM11, FF15]), sondern auch zu der sexuellen Orientierung einer Person [Ne15] hergestellt.

Während genetischen Daten in den meisten Gesetzgebungen ein besonderer Schutz zugesprochen wird, trifft dies jedoch nicht unbedingt auch auf epigenetische Daten zu. Der US Genetic Information Nondiscrimination Act (GINA) beispielsweise bezieht sich explizit auf *genetische* Daten und im wissenschaftlichen Sinn sind diese nicht gleichzusetzen mit *epigenetischen* Daten [RCM09, Dy15].

Unsere Datenschutzbedenken werden außerdem durch die Existenz von Datenbanken verstärkt, in denen verschiedenste Arten von biomedizinischen Daten – z.B. für Forschungszwecke – gesammelt und bereitgestellt werden. Oftmals veröffentlichen Forscher in solchen Datenbanken ihre biomedizinischen Studien. Auf diese Art sind eine Vielzahl epigenetischer Datensätze (in pseudonymisierter Form) bereits heute online für jedermann frei zugänglich. In Anbetracht des Milliarden-Dollar-Geschäfts mit dem Verkauf von privaten Gesundheitsdaten und Brokern, die solche pseudonymisierten Patientendaten aus verschiedenen Quellen untereinander verbinden [Yo, Th], bedarf es zweifelsohne der Möglichkeit, verbundene Privatsphärerisiken zu quantifizieren und abzuwehren.

Meine Dissertation [Be18a] betrachtet Verfahren zur Quantifizierung und Abwehr solcher Privatsphärerisiken. Neben dem Genom konzentrieren wir uns dabei auf zwei der wichtigsten gesundheitsrelevanten epigenetischen Elemente: microRNAs und DNA-Methylierung. Für die Quantifizierung der Privatsphärerisiken betrachten wir mehrere realistische Angriffsszenarien: (1) Verknüpfung von Profilen über die Zeit, Verknüpfung verschiedener Datentypen und verwandter Personen, (2) Feststellung der Studienteilnahme und (3) Inferenz von Attributen. Unsere Resultate bekräftigen, dass die Privatsphärerisiken solcher Daten ernst zu nehmen sind. Außerdem präsentieren und evaluieren wir Lösungen zum Schutz der Privatsphäre, sowohl auf Basis von Veränderung der Daten, als auch auf Basis von Kryptographie. Sie reichen von der Anwendung von Differential Privacy unter Berücksichtigung des Nutzwertes bis zu kryptographischen Protokollen zur Auswertung eines Random Forests. Während Differential Privacy besonders geeignet ist für das Veröffentlichen von Datensätzen und Statistiken, erlauben kryptographische Anwendungen die sichere Speicherung und Analyse von Daten ohne deren Qualität zu beeinflussen.

Die Dissertation basiert auf unseren vorangegangenen Arbeiten, welche allesamt auf Top-Tier-Konferenzen im Bereich der IT-Sicherheit präsentiert wurden [Ba16a, Ba16b, Ba17, Be18b] und welche sich als Leitmotiv durch die Dissertation ziehen.

2 Hintergrund

Bevor wir beginnen die Privatsphärerisiken, die mit epigenetischen Daten einhergehen, zu analysieren, geben wir zuerst eine kurze Einführung in die Thematik von Differential Privacy und erklären die Relevanz der Privatsphäre-Nutzen-Abwägung im Hinblick auf biomedizinische Daten und Anwendungsfälle. Im Anschluss fassen wir die wichtigsten Eigenschaften der drei von uns betrachteten Elemente zusammen: dem Genom, microRNAs und DNA-Methylierung.

2.1 Differential Privacy

Differential Privacy ist eine Technik, die darauf abzielt die Privatsphäre von Individuen zu schützen, die Teil einer Datenbank sind. Zu diesem Zweck werden die Inhalte der Datenbank unter Hinzufügen von Rauschen so verändert, dass die Zugehörigkeit einzelner Individuen zur Datenbank von einem Angreifer nicht mehr zweifelsfrei festgestellt werden kann. Gleichzeitig soll jedoch die statistische Aussagekraft der Datenbank und damit die Genauigkeit von statistischen Anfragen an die Datenbank maximiert werden. Da diese Technik die Daten verändert, entsteht allerdings ein Konflikt zwischen der Privatsphäre der Teilnehmer und dem Nutzwert der Daten. Der Nutzen der Daten ist dabei stark anwendungsabhängig. In vielen medizinischen Szenarien, wie beispielsweise einer Diagnose, ist der Nutzwert besonders kritisch. Er kann beispielsweise als die Genauigkeit der Diagnose definiert und gemessen werden.

2.2 Genetik und Epigenetik

Das Genom enthält das sogenannte Erbgut eines Organismus und verändert sich mit Ausnahme geringster Mutationen über die gesamte Lebensdauer nicht. Es ist Informationsträger und bestimmt die potentiellen Verhaltensmöglichkeiten einer Zelle. Die Genexpression, also die Teile des Genoms welche in einer Zelle aktiv sind, bestimmt das eigentliche Verhalten unserer Zellen. Beim Menschen wird das Genom durch eine Rekombination von den Eltern an ihre Kinder vererbt.

Die Epigenetik ist ein Forschungsgebiet, das sich mit solchen Faktoren befasst, welche die Aktivität und Entwicklung einer Zelle beeinflussen, jedoch nicht auf Veränderungen des Genoms zurückzuführen sind. Solche externen Faktoren sind beispielsweise Chemikalien in der Umgebung, Alterung oder auch Ernährung. Epigentik bezieht sich zudem auf die direkten Veränderungen in der Zelle (z.B. DNA-Methylierung), welche Einfluss auf die Genexpression nehmen, ohne das Genom selbst zu verändern. Im Gegensatz zum Genom bildet die Epigenetik einen Schnappschuss unseres aktuellen Zustands ab und verändert sich somit über die Zeit.

MicroRNAs (abgekürzt auch miRNAs) sind epigenetische Elemente, die ein wichtiger Bestandteil der Genregulation (der Regulation von Genexpression) sind. Studien haben gezeigt, dass die miRNA Expression – also das Vorhandensein bestimmter miRNAs –

in direktem Zusammenhang mit neurodegenerativen Erkrankungen, Herzerkrankungen, Diabetes und einer Vielzahl von Krebsarten steht [Lu05, Wo07, JB07, QM11, FF15].

DNA-Methylierung ist eines der am besten verstandenen epigenetischen Elemente. Es ist ein essentieller Regulator der Gentranskription (d.h. der Synthese von RNA anhand der DNA). Von der Norm abweichende DNA-Methylierungsmuster wurden so mit verschiedensten Krebsarten in Verbindung gebracht [EH02, DS04, Vo13]. Ein DNA-Methylierungsprofil beschreibt an welchen Stellen des Genoms eine Methylierung vorliegt.

3 Verknüpfbarkeit von miRNA Expressionsprofilen

Insbesondere aufgrund der zeitlichen Veränderbarkeit epigenetischer Daten untersuchen wir ein Angriffsszenario, in dem epigenetische Profile über die Zeit rückverfolgt und verknüpft werden. Im Speziellen betrachten wir die Verknüpfbarkeit von miRNA Expressionsprofilen. Wir zeigen auf, dass die Variabilität von miRNA Expressionsprofilen keineswegs ein Garant für natürlichen Datenschutz ist und legen somit, als eine der ersten Arbeiten im Bereich epigenetischer Privatsphärerisiken überhaupt, den Grundstein für weitere Forschung.

Wir analysieren die sogenannte *zeitliche Verknüpfbarkeit* von miRNA Expressionsprofilen, indem wir zwei Arten von Angriffen präsentieren und diese daraufhin sorgfältig evaluieren. Der erste Angriff ist ein Identifikationsangriff. Das bedeutet, dass der Angreifer ein miRNA Expressionsprofil seines Ziels gegeben hat und damit das korrespondierende Profil in einer Datenbank sucht, welche solche Profile von einem anderen Zeitpunkt beinhaltet. Der zweite Angriff ist ein Abgleichungsangriff – eine Verallgemeinerung des Identifikationsangriffs. In diesem Fall besitzt der Angreifer bereits Zugriff auf eine Datenbank von miRNA Expressionsprofilen und versucht korrespondierende Profile in einer anderen Datenbank (von einem anderen Zeitpunkt) zu finden. Wie bereits zuvor dargestellt sind solche Datenbanken heute schon Realität. Sowohl durch Forscher, die diese Datenbanken zur Veröffentlichung von Studiendaten verwenden, als auch durch das Hacken oder den Verkauf von Patientendaten.

Unsere Angriffe evaluieren wir auf öffentlich verfügbaren, pseudonymisierten Datensätzen von Verlaufsstudien aus dem Gene Expression Omnibus (GEO) [Ge]. In unseren Experimenten zeigen wir die Effektivität dieser Angriffe. So konnten wir beim Abgleichen von blut-basierten miRNA Expressionsprofilen, die in einem Abstand von einer Woche erfasst wurden, eine Erfolgsrate von bis zu 90% verzeichnen. Außerdem zeigen wir, dass eine Variation des Zeitraums zwischen den Profilen von einer Woche bis zu einem Jahr kaum Einfluss auf die Erfolgsrate des Angriffs hat.

Im Anschluss widmen wir uns dem Schutz vor solchen Angriffen und präsentieren zwei mögliche Abwehrmechanismen, welche die Daten verändern. Unser erster Abwehrmechanismus arbeitet durch gezieltes Verbergen einer Teilmenge der miRNA Expressionen (z.B. solche, die für einen gegebenen medizinischen Anwendungsfall irrelevant sind). Unser zweiter Abwehrmechanismus hingegen nutzt Differential Privacy, um die miRNA Expres-

sionsprofile zu verrauschen. Hierbei kann jeder Teilnehmer selbst den Grad der Verrauschung wählen – beispielsweise bei der Teilnahme an einer Studie.

Wir evaluieren unsere Abwehrmechanismen auf den gleichen Datensätzen, die wir bereits zur Evaluation der Angriffe genutzt haben. Dabei betrachten wir nicht nur die Auswirkungen der Abwehrmechanismen auf die Privatsphäre, sondern auch auf den biomedizinischen Nutzen der Daten. Die Privatsphäre messen wir invers proportional zum Angriffserfolg. Den Nutzwert der Daten schätzen wir anhand einer Klassifizierung der Profile in gesund oder erkrankt mit Hilfe von krankheitsspezifischen Studien. Dabei steht die Genauigkeit der Krankheitsprognose in einem Zielkonflikt mit dem Datenschutz. Unsere Experimente ergeben, dass die Methode auf Basis des Verrauschens diesen Zielkonflikt hier geeignet abwägen kann. Unter Anwendung von Differential Privacy ist es möglich die Verknüpfbarkeit um mindestens 50% zu senken, während die Genauigkeit der Prognose kaum beeinträchtigt ist ($< 1\%$).

4 Gruppenzugehörigkeit von miRNA Expressionsprofilen

Eine weitere Gruppe von Angriffen betrifft sogenannte *Gruppenzugehörigkeits-Angriffe*. Diese beziehen sich meist auf biomedizinische Studien, welche zusammengefasste Statistiken (wie z.B. den Durchschnitt) über die genutzten Daten veröffentlichen. Während diese Art von Angriffen bereits für genetische Daten untersucht wurde, sind wir die Ersten, die diese für epigenetische Daten erforschen. Bei einem Gruppenzugehörigkeits-Angriff auf miRNA Expressionsprofilen hat der Angreifer Zugang zu einem solchen Profil und versucht nun die Zugehörigkeit des Profils zu einer Gruppe von Personen anhand zusammengefasster Statistiken zu ermitteln. In anderen Worten: ein Angreifer kann so die Studienteilnahme seiner Zielperson feststellen und somit beispielsweise herausfinden, ob diese Person Teil einer Gruppe erkrankter Personen einer Krebsstudie ist.

Wir präsentieren zwei solche Angriffe: einen Angriff basierend auf der L_1 Distanz und einen Angriff basierend auf dem Likelihood-Quotienten-Test. Die Angriffe evaluieren wir auf Statistiken sowohl über krankheitsspezifischen als auch zufällig zusammengesetzten Gruppen. Unsere Ergebnisse demonstrieren, dass krankheitsspezifische Gruppen besonders anfällig für diese Angriffe sein können, mit einer Richtig-positiv-Rate von bis zu 77% bei einer Falsch-negativ-Rate von weniger als 1%. Außerdem zeigen wir, dass der auf dem Likelihood-Quotienten-Test basierende Angriff den größten Angriffserfolg zu verzeichnen hat und leiten eine theoretische Obergrenze für diesen Angriffserfolg her.

Wir stellen zwei Techniken vor, um epigenetische Daten vor solchen Angriffen zu schützen. Ähnlich zu der Verknüpfung von epigenetischen Daten, setzen wir hier wieder auf das Verbergen von Teildaten, sowie Differential Privacy als Alternative. Im Falle von Differential Privacy, betrachten wir außerdem zwei Angreifermodelle mit unterschiedlichen Annahmen über das Wissen des Angreifers. Wir evaluieren unsere Techniken wieder sowohl mit Hinblick auf die Privatsphäre als auch den Nutzwert der Daten. Hierbei kann der auf Differential Privacy basierende Abwehrmechanismus die Privatsphäre in miRNA basierten Studien ohne große Verluste beim Nutzen der Daten schützen, solange die Datensätze verhältnismäßig groß sind. Es zeigt sich, dass der Einfluss des Rauschens auf

den Nutzwert der Daten in solchen Statistiken wesentlich höher ist als bei anderen Arten von Datensätzen (wie beispielsweise solche mit einzelnen Expressionsprofilen). Unsere theoretische Herleitung zeigt, dass der Angriffserfolg sich linear in steigender Anzahl an Gruppenmitgliedern verschlechtert. In Kombination mit der aktuell bekannten Anzahl an miRNAs empfehlen wir daher zusammengesetzte Statistiken über miRNA Expressionsprofile lediglich für Datensätze von einigen hundert Individuen zu veröffentlichen.

5 Genotyp-Inferenz aus DNA-Methylierungsprofilen

Während sich unsere bisherigen Untersuchungen auf einzelne Typen biomedizinischer Daten beschränkten, ist es von ebenso immenser Bedeutung Abhängigkeiten zwischen mehreren Arten von Daten zu betrachten. Solche Abhängigkeiten ermöglichen es beispielsweise verschiedene Arten biomedizinischer Daten untereinander zu verknüpfen. Wir untersuchen daher die Abhängigkeiten zwischen dem Genom und dem epigenetischen Element der DNA-Methylierung. Dabei zeigen wir, dass die Veröffentlichung eines DNA-Methylierungsprofils einer Veröffentlichung des eigentlichen Genoms gleicht.

Im Speziellen zeigen wir, dass eine geringe Menge an vom Genom beeinflussten DNA-Methylierungsregionen ausreicht, um die entsprechenden Genotypen abzuleiten. Dies kann daraufhin ausgenutzt werden, um ein DNA-Methylierungsprofil auf das entsprechende Genom abzubilden. Wir formalisieren diese Art eines *Wiedererkennungsangriffs* und stellen einen statistischen Test bereit, der falsche Verknüpfungen erkennen und eliminieren kann.

Wir evaluieren diesen Wiedererkennungsangriff anhand eines großen Datensatzes von Genom- und DNA-Methylierungsdaten, welche von Mutter-Kind-Paaren erfasst wurden. Damit ist es uns auch möglich den Angriff für solche Fälle zu evaluieren, in denen sehr ähnliche Genome von verwandten Personen beteiligt sind. Unsere Ergebnisse zeigen, dass ein Wiedererkennungsangriff selbst dann noch mit 97,5% Genauigkeit möglich ist, wenn ein DNA-Methylierungsprofil mit dem entsprechenden Genom in einer großen Genomdatenbank von über 2500 Genomen verknüpft werden soll. Unser statistischer Test erlaubt es uns weiterhin die wenigen falsch verknüpften Profile zu eliminieren.

Da unsere Evaluation ein erhebliches Privatsphärerisiko in der Veröffentlichung von DNA-Methylierungsprofilen aufzeigt, untersuchen wir potentielle Schutzmaßnahmen für DNA-Methylierungsprofile. Hierzu betrachten wir ein Szenario aus der medizinischen Praxis, in dem Patientendaten erhoben und zur Krankheitsdiagnose analysiert werden. Während der eigene Arzt die Patientendaten erhebt, wird die eigentliche Analyse oftmals von Drittanbietern durchgeführt. In der medizinischen Praxis können DNA-Methylierungsprofile beispielsweise genutzt werden, um den genauen Typ eines Tumors festzustellen. Dies kann mit Hilfe eines Random Forest Klassifizierers geschehen, wie 2015 von Danielsson et al. demonstriert wurde [Da15].

Während bisherige Forschung bereits Protokolle zur kryptographischen, privaten Auswertung von Entscheidungsbäumen entwickelt hat [Bo15], gehen wir ein Stück weiter und präsentieren ein System zur privaten Auswertung von Random Forests, welches in dem oben beschriebenen, typischen Szenario eingesetzt werden kann. Unser Protokoll basiert

auf homomorpher Verschlüsselung und wir beweisen dessen Sicherheit unter Annahme eines ehrlichen, aber neugierigen Angreifers – eine häufige Annahme in diesem Bereich. Das vorgestellte Protokoll ist prinzipell nicht anwendungsspezifisch für epigenetische Daten, sondern kann für jede Klassifikation eines Random Forests verwendet werden. Wir evaluieren unser System anhand echter DNA-Methylierungsprofile und der Random Forest Instanz die von Danielsson et al. vorgestellt wurde. Wir zeigen, dass sich der Kosten-Overhead in einem für unser Szenario akzeptablen Bereich hält.

6 Privatsphärerisiken durch Interdependenz biomedizinischer Daten

Unsere bisherigen Resultate haben die Notwendigkeit zur Quantifizierung und zum Schutz der Privatsphäre im Umgang mit biomedizinischen Daten dargelegt. Außerdem haben wir die damit einhergehenden Risiken in Bezug auf spezialisierte Angriffe untersucht. In einer logischen Konsequenz generalisieren wir daher die betrachteten Szenarien und präsentieren eine Methodik, die ein allgemeines Framework zur Quantifizierung von Privatsphärerisiken durch die Interdependenz biomedizinischer Daten darstellt. Dies ist der erste Schritt in Richtung einer umfassenden Betrachtung und Quantifizierbarkeit solcher Privatsphärerisiken im Umgang mit biomedizinischen Daten im Allgemeinen.

Dazu schlagen wir ein Modell basierend auf Bayesschen Netzen vor, welches wir im konkreten Fall sowohl mit genetischen als auch epigenetischen Daten instantiieren. Außerdem umfasst unser Modell die Verwandtheitsbeziehung zwischen Müttern und ihren Kindern, sowie die zeitliche Variabilität von epigenetischen Daten. Wir führen einen generischen Algorithmus zum Lernen der Struktur eines Bayesschen Netzes ein, der vorhandenes Vorwissen mit Trainingsdaten kombiniert, und beweisen dessen Korrektheit. Dieser Algorithmus ist nicht auf unser Modell beschränkt, sondern kann für Bayessche Netze im Allgemeinen verwendet werden. Dieser Algorithmus erlaubt es uns beispielsweise bekannte Relationen in unseren Bayesschen Netzen festzusetzen – wie die Vererbungsregeln des Genoms zwischen Mutter und Kind – und lernt unbekannte Relationen aus vorhandenen Trainingsdaten.

Wir nutzen die konkrete Instantiierung unserer Methodik dann, um eine umfassende Quantifizierung der Privatsphäre vorzunehmen. Hierzu legen wir dem Modell den Datensatz zu Grunde, der bereits für die Genotyp-Inferenz genutzt wurde, nutzen jedoch auch den longitudinalen Aspekt dieses Datensatzes aus. Wir quantifizieren die Privatsphäre anhand etablierter Privatsphäre-Metriken wie der erwartete Abweichung vom korrekten Ergebnis oder der Entropie. Unser Modell erlaubt es uns beliebige Angriffsszenarien durchzuspielen, die Ausgangs-Wissenslage des Angreifers zu variieren und ebenso festzulegen welche Attribute der Angreifer voraussagen möchte. In unseren Experimenten zeigen wir die Flexibilität dieser Methodik auf und bestätigen die inhärenten Privatsphärerisiken durch die Interdepenz der Daten. Wir können spezielle Angriffe reproduzieren und demonstrieren, dass die Privatsphärerisiken ernst zu nehmen sind.

Neben der Inferenz von Attributen dient unser Modell außerdem auch als Grundbaustein für weitere Anwendungen zur Quantifizierung der Privatsphäre. So zeigen wir, wie un-

ser Modell genutzt werden kann, um einen Verknüpfungsangriff zu simulieren. Wir messen auf diese Weise den Erfolg eines Angreifers, der eine Art Mutterschaftstest auf Basis von DNA-Methylierungsprofilen durchführt. Gegeben die DNA-Methylierungsprofile von Müttern versucht der Angreifer die DNA-Methylierungsprofile der entsprechenden Kinder in einer Datenbank auszusondern. Dabei kann der Angreifer aufgrund unseres Bayesschen Netzwerk Modells indirekt auch Wissen über Abhängigkeiten zwischen DNA-Methylierung und Genom ausnutzen, ohne jemals Genomdaten seiner Ziele zu besitzen. Die Evaluierung dieses Verknüpfungsangriffs unterstreicht unsere Privatsphärebedenken mit einer Erfolgsrate von 95%.

7 Zusammenfassung

Während sich bisherige Arbeiten im Bereich der Privatsphäre biomedizinischer Daten hauptsächlich auf genetische Daten fokussiert haben, demonstrieren wir in unserer Arbeit die Notwendigkeit für Methoden zur Quantifizierung und zum Schutz der Privatsphäre in Bezug auf weitere biomedizinische Daten – wie z.B. epigenetischer Daten. Indem wir mehrere Angriffsszenarien beleuchten und geeignete Abwehrmechanismen vorschlagen, schaffen wir das Bewusstsein für die Bedeutung dieser Art von Forschung. Wir betrachten unter anderem Verknüpfungs-, Identifikations- und Inferenzangriffe auf Patientendaten. Zudem gehen wir den ersten Schritt in Richtung einer umfassenden Sicht auf solche Angriffe und präsentieren eine generalisierte Methodik zur Quantifizierung der Privatsphäre für untereinander abhängige Daten.

Während die Evaluation dieser Angriffe auch die Werkzeuge liefert, die damit einhergehenden Privatsphärerisiken besser einschätzen zu können, schlagen wir ebenso geeignete Gegenmaßnahmen vor, um das Risiko dieser Angriffe zu mindern. Damit lassen sich die Ergebnisse unserer Arbeiten in drei Kategorien einteilen: (1) Tools zur Quantifizierung der Privatsphärerisiken, (2) Gegenmaßnahmen, welche die Datensätze verändern und somit Einfluss auf den Nutzwert derer haben können, und (3) Gegenmaßnahmen basierend auf Kryptographie. Wir heben dabei hervor, dass es entscheidend ist solche Gegenmaßnahmen in enger Zusammenarbeit mit biomedizinischen Experten zu entwerfen. Nur so können deren Belange beachtet werden und damit Lösungen entwickelt werden, welche in praxisrelevanten Szenarien wirklich verwendet werden können. Insbesondere für Gegenmaßnahmen, welche die Datensätze verändern oder verrauschen, bedeutet dies die kritische Abwägung des Konflikts zwischen Privatsphäre, dem Nutzen der Daten und einer einfach nutzbaren Lösung. Lässt man diesen Konflikt außer Acht, wird man keine Gegenmaßnahme entwickeln können, welche in der Realität Anwendung findet. Im schlimmsten Fall kann das Missachten des Nutzwertes beispielsweise dazu führen, dass die Gesundheit von Patienten durch ungenaue Diagnosen gefährdet wird [Fr14]. So sind Maßnahmen basierend auf dem Verrauschen von Daten hauptsächlich für die Veröffentlichung von Datensätzen geeignet, während Diagnosemaßnahmen eher durch kryptographische Methoden abgesichert werden sollten. Im Falle der von uns vorgestellten Gegenmaßnahmen gehen wir jeweils auf diesen vorgestellten Trade-Off ein und evaluieren eine Vielzahl von Parametern, um eine geeignete Abwägung zu ermöglichen.

Literaturverzeichnis

[Ba16a] Backes, Michael; Berrang, Pascal; Hecksteden, Anne; Humbert, Mathias; Keller, Andreas; Meyer, Tim: Privacy in Epigenetics: Temporal Linkability of MicroRNA Expression Profiles. In: Proceedings of the 25th USENIX Security Symposium (Security). USENIX Association, S. 1223–1240, 2016.

[Ba16b] Backes, Michael; Berrang, Pascal; Humbert, Mathias; Manoharan, Praveen: Membership Privacy in MicroRNA-based Studies. In: Proceedings of the 23rd ACM Conference on Computer and Communication Security (CCS). ACM, S. 319–330, 2016.

[Ba17] Backes, Michael; Berrang, Pascal; Bieg, Matthias; Eils, Roland; Herrmann, Carl; Humbert, Mathias; Lehmann, Irina: Identifying Personal DNA Methylation Profiles by Genotype Inference. In: Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P). IEEE, S. 957–976, 2017.

[Be18a] Berrang, Pascal: Quantifying and Mitigating Privacy Risks in Biomedical Data. Dissertation, Saarland University, 2018.

[Be18b] Berrang, Pascal; Humbert, Mathias; Zhang, Yang; Lehmann, Irina; Eils, Roland; Backes, Michael: Dissecting Privacy Risks in Biomedical Data. In: Proceedings of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2018.

[Bo15] Bost, Raphael; Popa, Raluca Ada; Tu, Stephen; Goldwasser, Shafi: Machine Learning Classification over Encrypted Data. In: Proceedings of the 22nd Annual Network and Distributed System Security Symposium (NDSS). The Internet Society, 2015.

[Cl10] Cloud, John: Why Your DNA Isn't Your Destiny. Time Magazine, 6, 2010.

[Da15] Danielsson, Anna; Nemes, Szilárd; Tisell, Magnus; Lannering, Birgitta; Nordborg, Claes; Sabel, Magnus; Carén, Helena: MethPed: a DNA methylation classifier tool for the identification of pediatric brain tumor subtypes. Clinical Epigenetics, 7(1):1, 2015.

[DS04] Das, Partha M; Singal, Rakesh: DNA methylation and cancer. Journal of clinical oncology, 22(22):4632–4642, 2004.

[Dy15] Dyke, Stephanie OM; Cheung, Warren A; Joly, Yann; Ammerpohl, Ole; Lutsik, Pavlo; Rothstein, Mark A; Caron, Maxime; Busche, Stephan; Bourque, Guillaume; Rönnblom, Lars et al.: Epigenome data release: a participant-centered approach to privacy protection. Genome Biology, 16(1):1–12, 2015.

[EH02] Esteller, Manel; Herman, James G.: Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. The Journal of Pathology, 196(1):1–7, 2002.

[EN14] Erlich, Yaniv; Narayanan, Arvind: Routes for breaching and protecting genetic privacy. Nature Reviews Genetics, 15(6):409–421, 2014.

[FF15] Feinberg, Andrew P; Fallin, M Daniele: Epigenetics at the Crossroads of Genes and the Environment. JAMA, 314(11):1129–1130, 2015.

[Fr14] Fredrikson, Matthew; Lantz, Eric; Jha, Somesh; Lin, Simon; Page, David; Ristenpart, Thomas: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Security Symposium (Security). USENIX Association, S. 17–32, 2014.

[Ge] Gene Expression Omnibus. `http://www.ncbi.nlm.nih.gov/geo`. Accessed: 29/01/2019.

[Hu13] Humbert, Mathias; Ayday, Erman; Hubaux, Jean-Pierre; Telenti, Amalio: Addressing the concerns of the Lacks family: quantification of kin genomic privacy. In: Proceedings of the 20th ACM Conference on Computer and Communication Security (CCS). ACM, S. 1141–1152, 2013.

[JB07] Jones, Peter A; Baylin, Stephen B: The epigenomics of cancer. Cell, 128(4):683–692, 2007.

[Lu05] Lu, Jun; Getz, Gad; Miska, Eric A; Alvarez-Saavedra, Ezequiel; Lamb, Justin; Peck, David; Sweet-Cordero, Alejandro; Ebert, Benjamin L; Mak, Raymond H; Ferrando, Adolfo A et al.: MicroRNA expression profiles classify human cancers. Nature, 435(7043):834–838, 2005.

[Ne15] Ngun et al., Tuck: Abstract: A novel predictive model of sexual orientation using epigenetic markers. In: American Society of Human Genetics 2015 Annual Meeting. 2015.

[QM11] Qureshi, Irfan A; Mehler, Mark F: Advances in epigenetics and epigenomics for neurodegenerative diseases. Current neurology and neuroscience reports, 11(5):464–473, 2011.

[RCM09] Rothstein, Mark A; Cai, Yu; Marchant, Gary E: The ghost in our genes: legal and ethical implications of epigenetics. Health matrix (Cleveland, Ohio: 1991), 19(1):1, 2009.

[Th] The Black Market For Stolen Health Care Data. `http://www.npr.org/sections/alltechconsidered/2015/02/13/385901377/the-black-market-for-stolen-health-care-data`. Accessed: 29/01/2019.

[Vo13] Vogelstein, Bert; Papadopoulos, Nickolas; Velculescu, Victor E; Zhou, Shibin; Diaz, Luis A; Kinzler, Kenneth W: Cancer genome landscapes. Science, 339(6127):1546–1558, 2013.

[Wo07] Wood, Laura D; Parsons, D Williams; Jones, Siân; Lin, Jimmy; Sjöblom, Tobias; Leary, Rebecca J; Shen, Dong; Boca, Simina M; Barber, Thomas; Ptak, Janine et al.: The genomic landscapes of human breast and colorectal cancers. Science, 318(5853):1108–1113, 2007.

[Yo] Your private medical data is for sale – and it's driving a business worth billions. `https://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns`. Accessed: 29/01/2019.

Pascal Berrang wurde am 6. Dezember 1991 in Saarbrücken, Deutschland geboren. Er begann sein Informatikstudium 2010 an der Universität des Saarlandes und trat 2013 – direkt nach Abschluss seines Bachelorstudiums – der Graduiertenschule an der Universität des Saarlandes bei. Er promovierte dort am IT-Sicherheitsinstitut CISPA (*Center for IT Security, Privacy and Accountability*) in der Gruppe von Prof. Michael Backes mit einem Fokus auf die IT-Sicherheits und Datenschutz Aspekte von biomedizinischen Daten. Seine interdisziplinäre Forschung erfolgte unter anderem in Kollaboration mit Biomedizinern des DKFZ (*Deutsches Krebsforschungszentrum*) und resultierte in mehreren Top-Tier-Publikationen. Seine Promotion schloss er 2018 mit Auszeichnung ab.